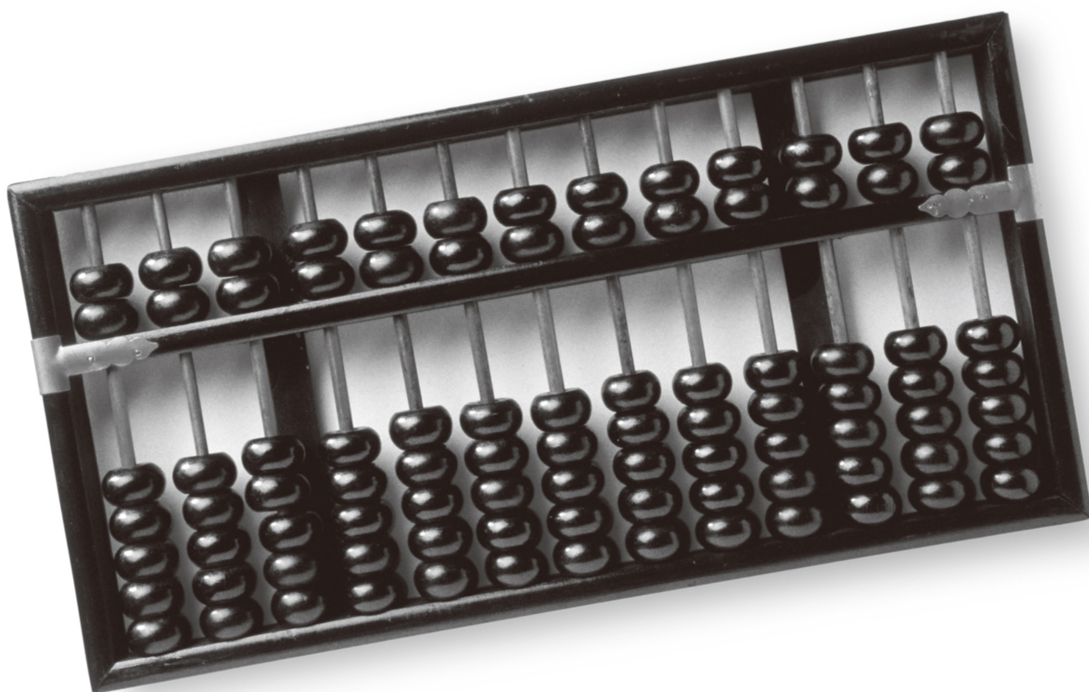
TURING

有趣的统计

75招学会数据分析

STATISTICS HACKS™



[美] Bruce Frey 著

邹澍 译

O'REILLY®

HACKS SERIES™

人民邮电出版社
POSTS & TELECOM PRESS

数字版权声明

作者介绍



Bruce Frey

博士，连环画收藏家和电影爱好者，现在堪萨斯大学教授研究生统计学课程，并且是该校教育心理与研究项目助理教授，在教学过程中屡获殊荣。Bruce获得的主要成就包括：青少年时期荣获堪萨斯州大富翁锦标赛第三名，大学时荣获堪萨斯州电影节第二名，中年时在堪萨斯州劳伦斯举办的得州扑克锦标赛中获得不错的第三名。

译者介绍



邹澍

热爱统计决策，热衷探究现象背后的统计原理。曾学习过统计学和心理学。目前是移动端用户体验工程师，密切关注用户体验前沿领域，在移动端应用需求研究和设计方面经验丰富。个人主页：uxoffer.com。

TURING

有趣的统计
75招学会数据分析

Statistics Hacks

[美] Bruce Frey 著
邹 澍 译

O'REILLY®

Beijing • Cambridge • Farnham • Köln • Sebastopol • Tokyo

O'Reilly Media, Inc.授权人民邮电出版社出版

人民邮电出版社
北 京

图书在版编目（ＣＩＰ）数据

有趣的统计 : 75招学会数据分析 / (美) 弗雷
(Frey,B.) 著 ; 邹澍译. -- 北京 : 人民邮电出版社,
2014.9
ISBN 978-7-115-35621-5

Ⅰ. ①有… Ⅱ. ①弗… ②邹… Ⅲ. ①统计数据－统
计分析（数学） Ⅳ. ①O212.1

中国版本图书馆CIP数据核字(2014)第101818号

内 容 提 要

概率在我们的日常生活中扮演着重要角色。每个人独特的基因组成，学生的在校成绩，赌博游戏中的下注，体育比赛的结果等，到处都有概率的身影。本书分为 6 章，共包括 75 个“黑客技艺”（Hack），揭示了如何通过推论统计学来理解事物的运作方式，发现变量之间的相关性，以及透过局部样本分析推断出总体特征，做出异常精准的预测。

本书适合从事数据统计、分析工作或对统计分析感兴趣的读者。

◆ 著 [美] Bruce Frey
译 邹 澍
责任编辑 李松峰
执行编辑 李 静 张 庆
责任印制 焦志炜

◆ 人民邮电出版社出版发行 北京市丰台区成寿寺路11号
邮编 100164 电子邮件 315@ptpress.com.cn
网址 http://www.ptpress.com.cn
北京 印刷

◆ 开本：800×1000 1/16
印张：18.25
字数：431千字 2014年 9 月第 1 版
印数：1－4 000册 2014年 9 月北京第 1 次印刷
著作权合同登记号 图字：01-2012-9300号

定价：59.00元

读者服务热线：(010)51095186转600 印装质量热线：(010)81055316

反盗版热线：(010)81055315

广告经营许可证：京崇工商广字第 0021 号

版 权 声 明

目　　录

作者中文版序

统计学作为一种解决问题的方法已经传授了几百年。在最初的几个世纪，统计学只是一套数学法则，用于确定某些事情发生的可能性。其中绝大部分法则都是具有数学天赋的人为了赌钱时稳赚不赔而想出来的。有时候，他们会与别人分享自己的发现，估计那时候他们的钱已经多得花不完，得考虑捐给慈善机构了。在最近大约 150 年，统计学的范畴不断扩充，已经包括了根据样本数据准确推断总体数据特征的方法。使用样本来描述整体属于推论统计学的范畴。推论统计学是统计学花园中绝美的花朵。

统计学的发展日新月异。目前统计学的主要应用是解决日常生活中的难题。市面上的统计学参考书虽然也致力于解决这些难题，比如推论统计学如何检验假设，如何回答研究问题，但很少探讨如何更有趣、更好玩地运用统计学。我希望通过本书改变这一现状。

本书由很多 Hack 组成。每个 Hack 都会用一种聪明的方法解决一个有意思的问题。这些问题涉及如何用统计学来回答日常难题、赢得比赛、赢到钱（向 500 年前的先辈们致敬），我会把自己能想到的所有方法毫无保留地教给你。此外，本书也涉及指导研究的传统统计学方法，你会看到有关检验和多次回归的 Hack，掌握了本书的技艺后，你就能在扑克牌游戏中获胜，破解秘密数据，比超人更加聪明。你还能学会自己解决遇到的问题。比如，“8 是你的幸运数字吗？”“裸婚能收获幸福的概率有多大？”设计你自己的 Hack 来揭开这些问题的谜底吧！

希望本书能给你带来愉悦的体验！祝你好运！

Bruce Frey 博士
2014 年 3 月

Preface for Chinese Edition of Statistics Hacks

Statistics has been taught as a way of solving problems for many 100's of years. For the first few centuries of the field, statistics was simply a set of mathematical rules for determining the likelihood of something happening. Most of those rules were developed by smart mathematicians who wanted to make

money gambling. Occasionally they would share their discoveries with others, perhaps when they were wealthy enough to give to charity. For the last 150 years or so, statistics has expanded to include methods of using a small sample of information to make fairly accurate guesses about a larger body of information. Using a sample to describe the larger population it represents is inferential statistics. And inferential statistics is the beauty of statistics.

Another more recent change is occurring in the field of statistics. The focus now is to apply statistics to everyday problems. While textbooks have long included many applied problems as examples of how inferential statistics can test hypotheses and answer research questions, few discuss the more interesting and fun ways to use statistics! I hope to change that with this book.

This book is full of different “hacks”. A hack is a clever way to solve an interesting problem. I’ve included all the ways I can think of to use statistics to answer everyday problems, to win games, and (in honor of our ancestors from 500 years ago) to win money. Oh, there are a few traditional discussions of statistical ways to conduct research and you’ll see hacks about t tests and multiple regression (whatever those are), but you’ll also find out how to win at poker, decode secret messages, and how to be smarter than Superman. And you’ll develop skills to answer your own interesting problems. Is 8 your lucky number? What are the chances that a marriage that started with a “naked” wedding will be successful? Find out by designing your own hacks!

I hope you’ll enjoy this book! Best wishes!

Bruce Frey, Ph.D. March, 2014

致　　谢

为本书做出贡献的人

下面这些人为本书贡献了自己的智慧，他们为我写作本书提供了素材和灵感。

- 约瑟夫 · 阿德勒（Joseph Adler）是 *Baseball Hacks*（O'Reilly 出版）一书的作者，在 VeriSign 公司的高级产品研发小组担任研究员，专注于用户认证、管理安全服务和 RFID 安全方面的问题。约瑟夫曾任 DoubleClick、美国捷运公司和 Dun & Bradstreet 等公司雇员和咨询顾问，他有着多年分析数据、构建统计模型和制定业务策略的经验。他毕业于麻省理工学院，获得计算机科学学士学位和计算机工程硕士学位。约瑟夫是全美职业棒球大联盟纽约洋基队的一位忠实球迷，但他也欣赏所有精彩的棒球比赛。约瑟夫和他的妻子住在硅谷，他们养了两只猫，家中装有 DirecTV 的卫星天线。
- 罗恩 · 黑尔–埃文斯（Ron Hale-Evans）是一名作家、思想家和游戏设计师。作为一名技术作家，他通过频繁的演出谋生。他拥有耶鲁大学心理学学士学位，大学期间还辅修了哲学。对思维的很多思考促使他创建了 Mentat Wiki（http://www.ludism.org/mentat），他也因此在最近写了《心理和脑与生活》（*Mind Performance Hacks*, O'Reilly 出版）。你可以在他的主页 http://ron.ludism.org 上找到他五花八门的（他自己就是这么写的）其他项目，包括他获奖的棋盘游戏、以及他的博客。罗恩的下一本书可能是关于游戏系统的，尤其是因为他给非常热爱但已停刊的《游戏杂志》（*The Games Journal*，http://www.thegamesjournal.com）所写的关于这个话题的一系列文章，在玩家和学者中一直非常受欢迎。如果你想通过电子邮件给罗恩发送一些容易上当受骗的出版商的名字，或者如果你只是想偷偷地知道他在想什么，你可以通过 rwhe@ludism.org 联系他（ludism 和 nudism 押韵，但和 Luddism 无关）。
- 布雷恩 · E. 汉森（Brain E.Hansen，27 岁，在得克萨斯州达拉斯地区长大。在西班牙为宗教使命服务两年后，他进入得克萨斯州农工大学学习并于 2004 年毕业，获石油工程理学学士学位。目前，他在一家总部位于得克萨斯州欧文市的大型独立石油和天然气勘探生产公司担任油藏工程师。

- 吉尔·H. 罗米尔（Jill H. Lohmeier）拥有马萨诸塞大学阿默斯特分校认知心理学博士学位。她目前在堪萨斯大学担任学校课程评价和研究小组的评估负责人。吉尔喜欢户外运动，尤其是跑步、徒步旅行以及和她的孩子们踢足球。
- 欧内斯特·E. 罗斯曼（Ernest E. Rothman）是罗得岛州纽波特沙尔瓦·瑞金纳大学（SRU）的数学科学系教授兼系主任。欧内斯特拥有布朗大学应用数学博士学位，在来沙尔瓦·瑞金纳大学（SRU）之前，他在纽约康奈尔理论中心任职。他的兴趣主要是科学计算、数学和统计学教育，以及 Mac OS X 的基础操作系统 Unix。你可以在 http://homepage.mac.com/samchops 上随时了解他的最新动态。
- 尼尔·J. 萨尔金德（Neil J. Salkind）曾经任教于堪萨斯大学，办公室在布鲁斯·弗雷对面，布鲁斯·弗雷享有统计黑客的声誉。尼尔除了是 *Statistics for Peoples Who(Think They)Hate Statistics*（*SAGE*）一书的作者，还是一名收集图书、会做饭、会修理老房子、驾驶沃尔沃 P1800 的发展心理学家，他还经常参加游泳大师赛。此外，他还写了 100 多本关于贸易的书和教材，并且经常与纽约的 StudioB 文学出版社合作。
- 威廉·斯科朗普斯基（William Skorupski）现任堪萨斯大学教育学院助理教授，教授心理测量学和统计学课程。2000 年，他从巴克内尔大学取得教育研究和心理学学士学位；2004 年，他在马萨诸塞大学阿默斯特分校取得心理测量方法博士学位。他的主要研究兴趣是将数学模型应用于数据的心理测量，包括使用贝叶斯统计来解决实际测量问题。他还喜欢将自己的统计学和概率知识应用于日常生活，比如和本书作者打扑克！

致谢

我要感谢所有为本书做出贡献的人，包括列入“为本书做出贡献者”部分的人和那些提供想法、审阅书稿的人，以及提供出处和资源建议的人。在这方面要特别感谢蒂姆·兰登（Tim Langdon），他给我送了一本哈里·布拉克史东（Harry Blackstone Jr）写的 *There's One Born Every Minute*（Jove 出版）平装本，这本书对本书中的很多技巧有极大的启发。

我要感谢我的编辑布莱恩·索耶（Brian Sawyer），是他果敢地推进着本项目的完成。他视野开阔，对什么是 Hack 什么不是 Hack 了然于胸。他大多数时候都是对的。（虽然不是每次都对：如何利用猴子选择肯塔基州大赛马的赢家本该作为 Hack 出现在本书里。也许下次可以将此收录进去吧……）布莱恩促成了这个项目的完成，尤其是在一连串不幸的骰子投掷，显示成功的几率看起来微乎其微的时候。

我要感谢最优秀的统计学作家尼尔·萨尔金德，感谢他对我的职业生涯和本书很多方面的帮助。

最重要的是，我要感谢我亲爱的妻子邦妮·约翰逊（Bonnie Johnson）。尽管在我最终交付本书的最后修订稿时，并不确定是否会想到她，但我想她将会在家里等着我。

前　言

无论你察觉与否，几率都在你生活中扮演着极其重要的角色。你个人独特的基因组成，在生命孕育之初即有些微的突变，而这些突变是基于特定的概率法则而发生的。你的在校成绩也会涉及某些人为误差——这可能是你自己造成的，也可能是其他人造成的，以致你的实际能力水平无法准确地反映在成绩报告单或利害攸关的测试中。职业生涯研究甚至表明，你的谋生之道可能并不是精心规划和准备的结果，而更有可能受到偶然事件的操控。当然，在几率游戏中你的命运完全取决于几率；几率对体育赛事的结果亦有着极其重要的影响。

幸好，我们现有一整套科学工具，即形形色色的统计应用程序，用以解决所谓的“命运影响体系”引起的诸多问题。推论统计学（inferential statistics）是一门完全基于概率本质的统计科学，不仅能让我们理解事物的运作方式，发现变量之间的相关性，还可透过局部样本分析推断出总体特征，做出异常精准的预测——没错，或许你已经想到，我们甚至偶尔可以通过推算出的概率，适当下注来赚些小钱。

本书集合了诸多统计技巧和应用工具。书中不仅囊括了统计学工具，还介绍了教育学、心理测量及实验研究设计等方面极具实用价值的工具，为社会学以及商务、游戏和博彩领域遇到的各种相关问题提供了解决方案。

倘若你是位顶尖科学家，睡梦中都在做统计运算，相信你会发觉此书趣味良多，它为你熟知的那些锈迹斑斑的旧工具找到许多充满创意的应用方法。倘若你只是在日常生活中喜欢科学探讨，以发掘奇绝的点子和巧解有趣的问题为乐，那你也大可放心：本书并非以纯学术式思维编写，你若觉得自己属于后一类人，那么这本书恰是你的不二之选。本书也不是专门写给统计学家看的，所以，哪怕你是统计学的门外汉，也照样能读出趣味。

另一方面，如果你选修了统计课程，或对学术性话题感兴趣，那么你会发现这本书是此类课程常用教科书的知音加伴侣，你的教科书与本书之间不存在任何相悖之处，了解一些貌似纯理论性的统计学工具在现实中的应用，并不会妨碍你的发展。事实上，你可以运用统计学去做许多非常酷的事情，这更像是一种娱乐而不仅是单纯的工作。

为什么称作Hack

Hacking 一词在传媒界声名狼藉。它一般用来指称那些以电脑作为武器，侵入或破坏他人电脑系统的人。然而，在专业程序员的圈子里，Hack 是指以“非常规的快捷方式”解决问题或者巧妙完成某件事的方案。在这一语境中，Hacker 一词便颇具赞赏之意，通常指代某些充满创造力，且拥有特定技能，能够出色完成任务的人。Hack 系列丛书[1]试图为 Hacker 正名，向正面意义上的 Hacker 行为致敬，向外行传递创造性参与的 Hacker 理念。要知道，观看他人如何动手操作系统并解决问题往往是学习一门新技术的捷径。

本书的技术核心是统计、测量和研究设计。计算机技术的发展向来与这些技术携手并进，因此，使用 Hack 来表述书中所要介绍的内容恰好与该词的本意完全相符。尽管书中只有一小部分内容涉及电脑 Hacking，但却介绍了大量巧妙而有实效的操作方法。

本书的组织结构

如果你愿意，可以从头至尾阅读本书。但鉴于书中所介绍的各种 Hack 自成一体，阅读时，你尽可随意浏览，翻到自己最感兴趣的部分去读。若想深入了解某项 Hack，还可循着书中的交叉引用跳转查阅。

前面的 Hack 更注重基础整合，通常会针对多种多样的问题提供广义的解决方案或战略方法。后面的 Hack 则更趋具体化，例如提供更具针对性的技巧来帮助我们赢得游戏，或者单纯提供信息，让我们充分认识到自己身边的情况。

全书按照不同的主题，分为以下几章。

第 1 章　基础知识

这部分 Hack 可以作为一个强大的基础工具合集，在你运用统计学 Hacking 解决麻烦时会频繁地用到它们。不妨将其想象成一套基础工具：它们是你手边的锤子、锯子以及不同规格的螺丝刀。

第 2 章　发现相关性

本章涵盖了用于发现、描述和测试变量相关性的多种统计方法。通过这些 Hack，你能化不可见为可见。

第 3 章　测量世界

这里为你呈现了测量身边世界的大量窍门和方法。你将学会如何正确提问，准确估算，甚至

注1：The Hack Series 是一套系列丛书，除本书外，还包括 *Baseball Hacks*、*Access Hacks*、*Mind Hacks*、*Excel Hacks* 等几部作品。——译者注

能够提高你在关键考试中的分数。

第 4 章　逆境制胜

本章是写给赌场玩家的。善用概率知识，可在得州扑克游戏中做出正确决定。这也同样适用于其他由概率定胜负的游戏。

第 5 章　游戏技巧

从“赢得大富翁”有奖电视游戏节目到只着眼于娱乐的体育赛事，本章为你呈现形形色色的 Hack，让你成为游戏里的最大赢家。

第 6 章　精明思考

本章可能是所有章节中最耗脑力的一章。理清思路，让我们来做脑力游戏，探索发现，使用本章的 Hack 揭开周围世界的神秘面纱。

本书排版约定

本书使用的排版规范如下所示。

- 楷体

  用于表示新的术语和概念。

- `等宽字体`

  用于表示 Excel 函数和代码示例。

- *`等宽斜体`*

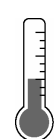
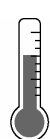
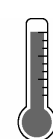
  用于表示用户需要根据自己提供的值进行更换的部分。

用以下图标标示的文字部分读者要特别注意。



该图标表示提示、建议或者一般注解。这部分内容是对当前主题的补充，很有用。

下面三个温度计图标出现在每个 Hack 的开始部分，说明这一主题的难易程度。

 初级　 中级　 专家级

Safari®Books Online



Safari Books Online（http://www.safaribooksonline.com）是应需而变的数字图书馆。它同时以图书和视频的形式出版世界顶级技术和商务作家的专业作品。

Safari Books Online 是技术专家、软件开发人员、Web 设计师、商务人士和创意人士开展调研、解决问题、学习和认证培训的第一手资料。

对于组织团体、政府机构和个人，Safari Books Online 提供各种产品组合和灵活的定价策略。用户可通过一个功能完备的数据库检索系统访问 O'Reilly Media、Prentice Hall Professional、Addison-Wesley Professional、Microsoft Press、Sams、Que、Peachpit Press、Focal Press、Cisco Press、John Wiley & Sons、Syngress、Morgan Kaufmann、IBM Redbooks、Packt、Adobe Press、FT Press、Apress、Manning、New Riders、McGraw-Hill、Jones & Bartlett、Course Technology 以及其他几十家出版社的上千种图书、培训视频和正式出版之前的书稿。要了解 Safari Books Online 的更多信息，我们网上见。

联系我们

请把对本书的评价和发现的问题发给出版社。

美国：

O'Reilly Media, Inc.
1005 Gravenstein Highway North
Sebastopol, CA 95472

中国：

北京市西城区西直门南大街 2 号成铭大厦 C 座 807 室（100035）
奥莱利技术咨询（北京）有限公司

O'Reilly 的每一本书都有专属网页，你可以在那里找到本书的相关信息，包括勘误表、示例代码以及其他信息。本书的网址是：http://www.oreilly.com/catalog/statisticshks

对于本书的评论和技术性问题，请发送电子邮件到：

bookquestions@oreilly.com

要了解更多 O'Reilly 图书、培训课程、会议和新闻的信息，请访问以下网站：

http://www.oreilly.com

我们在 Facebook 的地址如下：http://facebook.com/oreilly

请关注我们的 Twitter 动态：http://twitter.com/oreillymedia

我们的 YouTube 视频地址如下：http://www.youtube.com/oreillymedia

参与进来

要了解 Hacks 系列图书或者有意撰写该系列图书，请访问以下网站：

http://hacks.oreilly.com

O’Reilly Media, Inc.介绍

O’Reilly Media 通过图书、杂志、在线服务、调查研究和会议等方式传播创新知识。自 1978 年开始，O’Reilly 一直都是前沿发展的见证者和推动者。超级极客们正在开创着未来，而我们关注真正重要的技术趋势——通过放大那些“细微的信号”来刺激社会对新科技的应用。作为技术社区中活跃的参与者，O’Reilly 的发展充满了对创新的倡导、创造和发扬光大。

O’Reilly 为软件开发人员带来革命性的“动物书”；创建第一个商业网站（GNN）；组织了影响深远的开放源代码峰会，以至于开源软件运动以此命名；创立了 Make 杂志，从而成为 DIY 革命的主要先锋；公司一如既往地通过多种形式缔结信息与人的纽带。O’Reilly 的会议和峰会集聚了众多超级极客和高瞻远瞩的商业领袖，共同描绘出开创新产业的革命性思想。作为技术人士获取信息的选择，O’Reilly 现在还将先锋专家的知识传递给普通的计算机用户。无论是通过书籍出版，在线服务或者面授课程，每一项 O’Reilly 的产品都反映了公司不可动摇的理念——信息是激发创新的力量。

业界评论

“O’Reilly Radar 博客有口皆碑。”

——*Wired*

“O’Reilly 凭借一系列（真希望当初我也想到了）非凡想法建立了数百万美元的业务。”

——*Business 2.0*

“O’Reilly Conference 是聚集关键思想领袖的绝对典范。”

——*CRN*

“一本 O’Reilly 的书就代表一个有用、有前途、需要学习的主题。”

——*Irish Times*

“Tim 是位特立独行的商人，他不光放眼于最长远、最广阔的视野并且切实地按照 Yogi Berra 的建议去做了：‘如果你在路上遇到岔路口，走小路（岔路）。’回顾过去 Tim 似乎每一次都选择了小路，而且有几次都是一闪即逝的机会，尽管大路也不错。”

——*Linux Journal*

第 1 章

基础知识

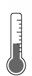
（Hack #1～#10）

统计学家用来探索世界、回答问题和解决难题的工具其实并不多，关键在于他们利用概率或者正态分布知识的方法，让他们能在千变万化的情境中解决问题。本章将为你介绍这些基本的Hack。

将已知的分布信息表述成概率[Hack #1]，这是统计黑客常用的基本技艺，与之类似的还有用小样本数据来准确描述数量较大总体中个体的分值[Hack #2]。懂得计算概率[Hack #3]的基本规则很重要。此外，如果你想基于统计作决策[Hack #4和Hack #8]，必须通晓显著性检验。

使估计中的错误[Hack #5]和得分中的错误[Hack #6]最小化，正确地解释数据[Hack #7]，是在不同情境下做到事半功倍的核心策略。成功的统计黑客能够轻而易举地识别出任何有组织观测的结果或实验操作的真正含义[Hack #9和Hack #10]。

弄懂这些核心工具的用法，学习和掌握后面的Hack将变得轻而易举。



HACK #1

1.1　不可不知的秘密

统计学家怎么让自己看起来比其他人聪明？

统计学作为一种科学方法，主要目的是对分数样本做概率解释。在深入学习前，需要简单了解一些术语，以便理解这个Hack，也为理解其他Hack打下基础。

样本是你目前收集到，就摆在你眼前的数值，用来表示既不在你眼前也没收集到的更大的数值总体。因为这些值几乎总是用来表示某一特征的存在或程度的数字，所以测量界把这些值称为分数。概率解释是对某件事情发生的可能性的解释。

概率是统计学的核心和灵魂。实际上，对统计学家的一种普遍看法就是，他们主要计算某些备受关注的事情（比如中彩票或是被雷击）发生的精确可能性。经验告诉我们，有办法计算骰子游戏结果可能性的人，同样有办法用为数不多的汇总统计数据来描述一大群人。

所以，通常统计学的教学中至少会花点时间来讲述概率的原理：计算不同组合的出现概率或者各种可能结果排列的方法。但是，统计学中更为常见的是*描述性统计*或是*推断性统计*，前者用以描述分数群组，后者仅用样本中包含的少量信息对分数总体进行估计。在社会科学中，“分数”常被用来描述人或是发生在人身上的事件。

当然，研究人员和测量人员（现实生活中最有可能使用统计的人）不局限于计算某种组合和排列的可能性。他们根本不需要计算连续3次扔一对骰子得到7的概率，他们能够运用不同的统计程序来回答复杂程度不同的问题。



如果你刚刚拿起骰子，那么这个概率是0.005[1]或是1%的1/2。如果你已经扔到了两个7，就有16.6%的概率扔到第三个7。

1.1.1　秘密

概率对统计学家的工作如此重要，关键原因是他们喜欢对实际或理论分布的分数进行概率解释。



分数的**分布**会列出一系列不同的值，有些情况下，还会给出每个值的数量。

比如，假设你知道刚刚参加的一次测验中，分数分布是25%的人得到10分，那我可能会说，我不需要认识你，也不需要了解与你有关的任何情况，就能知道你有25%的可能性得到10分。我同样可以说，你有75%的几率不得10分。我所做的只是获得关于某些值分布的已知信息，将其表述成概率。这是一种技巧，是所有统计学家都知道的秘密。实际上，这几乎是统计学家们所做的一切！

统计学家获取关于某些值分布的已知信息并将其表述成概率。我们有必要再次强调这句话（严格来说，这是第三遍）：*统计学家获取关于值分布的已知信息并将其表述成概率*。

天啊！这我们都能做到！这有何难？假设有一个空咖啡罐，里面有三个弹珠。再假设你知道其中只有一个弹珠是蓝色的。现在分布信息包含三个值：一个蓝色弹珠还有另外两个其他颜色的弹珠，这三个值构成了一个样本。三个弹珠里有一个是蓝色。噢，统计学家，闭上你的眼睛，请问我首次取出蓝色弹珠的几率是多少？1/3。33%。

注1：准确地讲应该是$1/(6 \times 6 \times 6)$，0.00462963，约等于0.005。——译者注

说句公道话，统计学家最常用的值及其分布通常比刚才那个从咖啡罐里取弹珠的场景稍微复杂或抽象一些，所以统计学家的工作看起来不是那么浅显易懂。比如，应用社会科学领域的研究人员总是用“值”来表示不同群体平均分之间的差异，或者两个或两个以上分数集的关系度量。其内在过程和刚才所举的咖啡罐的例子并无不同，不过是参考已知的值分布信息，做出对值的概率解释。

当然，关键是怎样才能知道这些奇特的、让统计学家感兴趣的值的分布？怎样才能了解平均差的分布或两组变量间关系程度的分布？方便的是，研究人员和数学家前辈已经发明或发现了多种公式和定理、经验法则、思想体系和假说，让我们了解到研究者最常用到的复杂值的分布情况。这些工作前人已经为我们做好了。

1.1.2 不太光彩的小秘密

统计学家获得分数分布的已知信息、将其表述为概率的方法中，多数必须满足一些前提，才能够确保概率解释的准确。其中一个几乎永恒不变的必要前提就是：样本值必须从分布中*随机抽取*。

请注意，在叙述咖啡罐的例子时，我插了一句“闭上你的眼睛”。如果抽样过程不是随机的，而是被其他一些因素所引导，那么得出的相应概率就错了；最糟糕的是，我们无法了解错误的程度。现今，也许绝大部分应用心理学和教育学研究都不是随机采样的。

比如，选修《心理学导论》课程的大学生，构成了很多心理学研究的样本。由于贪图便利，教育研究人员常用自家附近的小学生充作样本。这是社会科学研究者常常容忍、忽略或担心的一个问题，但不管怎样，非随机抽样是很多社会科学研究中存在的一个局限。



1.2 仅用两个数字描述世界

本书介绍的大部分统计解决方案和工具之所以行之有效，只因为你能通过样本对总体进行精确推断。获得以上推断技巧需要用到的元工具、主要指导方针和所有秘密之最就是中心极限定理。

每当你试图描述一组分数时，统计学都能为你提供解决方案。有时你想描述的整组分数全都摆在眼前，这时完成该任务的方法称作*描述性统计*。更常见的情形是，你只能看到欲描述的一组分数中的一部分，但仍想描述整个组。这种概括性方法称作*推断性统计*。在推断性统计中，你想要推断的整个分数群组叫做*总体*，其中能看到的那一部分叫做*样本*。

从定义上看，不经直接观察就能有把握地描述由多个值构成的总体，想来颇似一种奇妙的把戏。然而，你只消运用三条信息——两个样本值和一个总体分数分布形态的假设，便可自信而准确地描述那不可见的总体，其结果准得令人称奇。这样一套推断程序就是所谓的*中心极限定理*。

1.2.1　统计学基础一点通

推断性统计用两个值来描述总体：平均数和标准差。

1. 平均数

若要描述值的样本，报告一组分数的合理概要比展示每个分数更高效。这个数值应该能够代表群组中所有分数以及它们的共性。因此，这个数值被视为一组分数的趋中趋势。

由于种种原因，通常情况下对趋中趋势的最佳度量是平均数[Hack #21]。平均数是所有分数的算术平均，即把群组中所有值相加求和并除以群组中值的数量。相比其他趋中度量（比如中位数、众数等），平均数可以提供关于群组分数的更多信息。

实际上，从数学上看，平均数具有一个有趣的属性。其计算方式（所有分数相加并除以分数的数量）所导致的副效应就是产生了一个和其他所有分数尽可能接近的数字。这个平均数会和群组中的一些分数比较接近，和另一些分数距离较远。但是如果你将这些距离相加，得到的总数是最小的。其他任何数字，无论是真实的还是想象的，与群组中各个分数的距离总和都不会比它更小。

2. 标准差

仅仅知道某一分布的平均值还不够，我们还需要知道有关分数变异性的信息。是多数接近平均数还是多数远离平均数？两个非常不同的分布可能有着相同的平均数，但变异度却大为不同。最常用的变异度量概括了每个分数和平均数的距离。

像平均数一样，承载更多信息的变异度量能用到分布内的所有数值。标准差就是这样一种变异度量。标准差是每个分数和平均数的平均距离。它统计某一分布中所有的距离并算出平均值：这里的“距离”是每个分数和平均数的距离。



另一个经常用于概括分布变异性的值是**方差**。方差是标准差的平方，在描述单一分布时并不是特别有用，但对比较不同分布的差异性很有帮助。方差常用作统计运算值，比如独立t检验[Hack #17]。

标准差公式看起来复杂得超出必要，但求和距离（当平均数被当做分割点时，负向距离总会抵消正向距离）在数学上的确有些复杂。故而有以下的方程式：

$$\sqrt{\frac{\sum(x-\text{平均数})^2}{n-1}}$$

其中Σ表示求和。x表示每个分数，n表示分数的数量。

1.2.2　中心极限定理

中心极限定理非常简单，但非常强大。该定理表述如下：

> 如果你从总体中随机抽取多个样本，那么每一样本的平均数趋于正态分布。

由此定理衍生出一系列的数学规则，用以准确估计上述虚构的样本平均数分布的描述值。

- 这些平均数的平均数（念起来真拗口）等于总体的平均数。凭借单样本的平均数，可对平均数的平均数做出很好的估计。
- 这些平均数的标准差等于样本标准差除以样本量的平方根（以字母n表示）：

$$\frac{\delta}{\sqrt{n}}$$

样本集合内的样本数量越大，这些数学规则产生的结果越准确，分布也更接近于正态曲线。



当样本数为30或30以上时，应用中心极限定理似乎足以得出准确结果。

1.2.3　那又如何

好吧，中心极限定理看起来有那么一点儿智力趣味性，并且毫无疑问能让统计学家们兴奋不已，但那又意味着什么呢？怎样才能用它来做点酷酷的事？

正如1.1节[Hack #1]中讨论的，这个所有统计学家都知道的有效秘诀是：获取一些值分布的已知信息，并将其表述为概率解释。当然，关键是怎样才能知道引起统计学家兴趣的不同类型值的分布？又该如何得知平均差异的分布或是两组变量关系大小的分布？答案是：中心极限定理。

比如，为了估计任何两个群组在某个变量上出现一定差异的概率，我们需要知道样本对应的总体平均数的分布。而当总体平均数不可见，甚至只是理论存在时，如何能够了解分布的状态？小伙子，答案就是中心极限定理！当样本可能从无限可能相关性的总体中抽取时，如何能够知道相关性（衡量两个变量间相关强度的指标）的分布？听说过中心极限定理吗，老兄？

既已知道正态曲线上值的比例[Hack #23]，中心极限定理又告诉我们这些概括性的值为正态分布，因此我能对每个统计结果标出概率。我能在我的结论和决策中用这些概率表示统计显著性水平（置信水平）。如果没有中心极限定理，我几乎无法做出任何关于统计显著性的解释。那将是何等乏味而悲哀的生活。

1.2.4 中心极限定理的实际应用

该定理在实际应用中，只需从总体中随机抽取几个样本值。例如，假设我手下新增了8个童子军。我的职责是教会他们打绳结。我猜在我指导过的童子军学员当中，这一批孩子并不是最聪明的。

在开口要求增加学费之前，我想要判断他们是否真的有点笨。我想知道他们的智商。我知道童子军的总体平均智商是100，但我注意到这8个童子军学员里没有一个智商超过100的。按理说，总该有个别超出这一水平的。这一组人是从平均总体里刻意选出的吗？也许，只是我的样本有点不同，并不代表所有童子军？如果使用中心极限定理的统计方法，就会提问道：

> 这个样本所代表的总体平均IQ可能是100吗？

如果我想知道我这组童子军是从什么样的整体中抽出来的，可以使用中心极限定理相当准确地估计总体的平均IQ和总体的标准差。我同样可以计算出样本平均IQ和总体平均IQ有多大差异。

我需要从手下的童子军那里获得一些数据以便进行以上计算。表1-1提供了一些不错的信息。

表1-1：童子军聪明程度

童子军	IQ
吉米	100
佩里	95
克拉克	90
莱克斯	92
尼尔	85
比利	88
格雷格	93
约翰	91

这8个IQ分数样本的描述性统计是：

- 平均IQ=91.75
- 标准差=4.53

于是我知道在我的样本组中，大部分个体的IQ分数在91.75的$4\frac{1}{2}$左右。不过，我更感兴趣的是他们所来自的那个未知的总体。利用中心极限定理我能够估计这一总体的平均数、标准差，更重要的是，能估计样本平均数在多大程度上偏离总体平均数。

- 平均IQ

我们的样本平均数可作为最好的估量依据，所以总体平均数很可能接近91.75。

- 总体中IQ分数的标准差

计算样本标准差的公式是专为估计总体标准差而设计的，所以推测总体标准差是4.53。

- 平均数的标准差

这才是真正关注的值。我们知道样本的平均数小于100，但那可能是偶然的吗？当从总体中随机抽取这含有8个数的样本时，样本的平均数会在多大程度上偏离总体平均数？这里要用到之前提过的方程式。输入样本值计算平均数的标准差，这通常称为平均数的标准误差：

$$\frac{\sigma}{\sqrt{n}} = \frac{4.53}{\sqrt{8}} = \frac{4.53}{2.83} = 1.60$$

由于中心极限定理，我们现在知道，8个童子军中大多数样本的平均数是在总体平均数 ± 1.6个IQ点的范围内。所以，这个平均数为91.75的样本不太可能是从平均数为100的总体中抽取出来的。总体平均数为93或者94，但不是100。

因为我们知道这些平均数是正态分布的，所以可以利用关于正态分布形态的知识[Hack #23]来生成一个精确的概率，即从平均数为100的总体中抽取出平均数为91.75的样本的概率。这种情况发生的概率低于1/100 000。看来我手下这批学习打绳结的孩子要比普通人难教一些。我也许可以多收一点学费。

1.2.5　其他生效领域

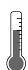
中心极限定理的一个模糊版本指出：

> 受很多随机作用和无关事件影响的数据最终呈正态分布。

因为这几乎适用于我们度量的所有事物，所以可以应用正态分布特征对多数可见和不可见概念做概率解释。

至此，我们还没有说到中心极限定理最厉害一条的推论：无论总体分布形态如何，从总体中随机抽取的平均数均呈正态分布。好好想想。即便你从中抽取样本的总体不是正态分布的，甚至走到了正态的反面（就像我的叔叔弗兰克那样），样本的平均数仍会是正态分布。

这是自然界相当了不起和便利的特征。不管我描述的总体是正态还是非正态、在地球上还是在火星上，这一要诀始终有效。



HACK #3

1.3　计算概率

我会中彩票吗？我会在一天内被雷击中又被公交车撞到吗？我所在的棒球队会在NCAA锦标赛中提前遇到令我们头疼的对手吗？统计学的核心要点就是判断事件发

生的可能性，并回答诸如此类的问题。计算概率的基本规则令统计学家有能力预测未来。

本书充满了有趣的难题，都可以通过绝妙的统计技巧解决。这些Hack中展示的方法在不同情境中以不同方式运用，同时，这些聪明解决方案中使用的很多程序能够起作用，是因为一个核心的元素：概率定律。

上述定律是一组简单、确定的关键原理，表明概率如何起作用，以及应当如何计算。以下两个基本定律可被视为一套基础入门工具，就像锤子和螺丝刀一样，大概足以解决大多数问题。

- 加法定律

几个互斥事件中任何一个发生的概率是各个事件发生的概率之和。

- 相乘定律

一系列独立事件都发生的概率是每个独立事件概率的乘积。

有了这两个工具，就足以回答日常生活中大部分关于“几率是多少”的问题。

1.3.1　关于未来的问题

当一个统计学家说出“1/10的可能性”这类话时，他就是对未来进行了一次预测。这或许是对一系列永远都无法检验的事件所做的假设性陈述，或许是对即将发生的事件不掺半点水分的如实解说。不管是哪种，他都是在对可能的结果进行统计学解释，所有的统计学家所说的话都无非如此[Hack #1]。



如果你能够理解以下表述，那么你就具有了像统计黑客一样行动和思考的必备能力：“如果有10件事情可能要发生，并且这10件事情发生的可能性相等，那么这10件事中任何一件发生的几率是1/10。”

科学研究中充满了可用统计来回答的问题，当然，还有概率定律的运用，但在实验室之外还有很多难题，比愚笨陈旧的科学问题更加重要的问题，比如骰子游戏。假设你是一名业余赌徒，家里的小孩想要双新鞋子。你下次掷出一对骰子的值会决定你的未来。那你也许想知道骰子扔出各种结果的可能性，而且是非常准确地知道这种可能性！

只凭这两件概率工具，就能回答你可能问到的三类最重要的概率问题。你提出的问题很可能是以下三种类型之一。

- ❑ 下一步出现某个特定结果的可能性是多少？比如，下面会掷出一个7吗？
- ❑ 下一步出现某组结果的可能性是多少？比如，下面会出现7或11吗？

- 下一步出现一系列结果的可能性是多少？比如，一对没被动过手脚的骰子真的能够整晚都不出现7吗（我说的是永远都不出现）？我的意思是，那真的可能吗？可能吗?!

概率术语

在谈论概率以及如何计算概率前，我们需要学会如何像统计学家一样说话。记得之前的“1/10的可能性”这句话吗？针对“几率是多少”的问题，共有三种回答方式。

用百分比来表示

1/10可以表述成10%。

用概率来表示

在可能性为1/10的情况下，成功概率就是1比9，即9分输1分赢。

用比例来表示

10%可以表述成0.10。从技术上讲，概率就该以比例来表述，否则就应当改用其他的名称。

1.3.2 特定结果发生的可能性

若你对某件事发生的可能性感兴趣，那么这里的“某件事”可以叫做获胜事件（在游戏情境中），或者只是一个你关注的结果（游戏以外的情境）。概率中的主要原则是用所关注的结果数除以全部结果的总数。全部结果的总数有时用大写的S表示（英文字母Set的首字母，代表集合），各种关注结果都用大写的A表示。（我猜这可能是因为A是字母表里的首字母，我是谁，数学家吗?）

于是有以下的概率基本公式：

$$\frac{A}{S}$$

计算任何特定结果或事件的几率，就是要算出这些结果的数量，并算出所有可能的结果数量，然后对两者进行比较。如果可能的结果为数很少，或者对获胜结果的描述很简单，仅包含单一事件，那么上述方法大抵很容易操作。

要回答一个典型的扔骰子问题，我们可以通过计算出两枚骰子点数之和等于期望数值的组合数量，来计算下次投掷时出现任何特定值的几率。然后，用那个数除以所有可能结果的总数。两个六面骰子，总共有36种可能的投掷结果。

比如，共有六种方式掷出7（我提前偷看了表1-2），6/36=0.167，所以任意一次投掷中掷出7的几率约为17%。



通过把每个骰子的总面数相乘，能够计算出可能投掷结果的总数：6×6=36。

1.3.3　出现一组结果的可能性

如果你对一组特定结果发生的可能性感兴趣，但并不关心具体发生的是哪一个，那么按照加法定律，可以把所有个体概率相加来计算总概率。为了回答我们的骰子问题，表1-2从“玩骰子行大运”[Hack #43]当中引用了一些信息，以便用比例表示掷出各种结果的几率。

表1-2：独立骰子投掷概率表

骰子投掷得数	结果的数量	概率
2	1	0.028
3	2	0.056
4	3	0.083
5	4	0.111
6	5	0.139
7	6	0.167
8	5	0.139
9	4	0.111
10	3	0.083
11	2	0.056
12	1	0.028
总数	36	1.0

表1-2提供了关于不同结果的信息。比如，有两种不同的方式掷出3[2]。两个获胜结果除以所有可能的结果总数36，得到0.056这个比例。所以，用两个骰子掷一次，大约有6%的几率掷出3。同时也请注意，所有可能事件的概率之和正好为1.0。

假设我们必须掷出几种结果中的一种，才能在赌局中获胜，那么让我们运用加法定律来查看这种情况下的获胜几率。比方说，只要你掷出了10、11、12中的任意一个就能赢，那么我们将这三个独立的概率相加：

$$0.083+0.056+0.028=0.167$$

你将有大约17%的概率掷出10、11或者12中任何一个。此处运用了加法定律，因为你关注于几个独立事件中的任意一件能否发生。

1.3.4　一系列结果发生的可能性

当概率问题变为“是否有若干件事情发生”，又将如何？当你想知道一连串特定事件是否发

注2：1+2或者2+1。——译者注

生时，这个问题总是被问到。事件发生的顺序通常不重要。

我们依然使用表1-2中的数据，以及之前例子中的三个值（10、11、12），就能够计算特定事件序列发生的几率。在给定连掷三次骰子的情况下，你连续掷得10、11、12的概率是多少？基于乘法定律，可将这三个独立概率相乘：

$$0.083 \times 0.056 \times 0.028 = 0.00013$$

这个非常特定的结果不太可能发生，其概率低于1‰，或者说1%的1/10。此处用到乘法定律，因为你感兴趣的是几个独立事件是否都会发生。

1.3.5　概率意味着什么

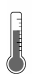
就本节介绍的Hack而言，概率即某事发生的可能性。我已将讨论限定在分析可能结果的背景下，这是思考概率的一个恰当方法。许多哲学家和社会科学家花费很多时间思考各种概念，诸如几率、未来和午饭该吃什么，在他们中间对概率有两种不同的视角。

分析视角。这是认识概率的经典视角，也是数学家和本条Hack所用方法的视角。分析视角识别所有可能的结果并计算获胜结果占所有结果的比例。这一比例就是概率。

我们通过概率解释来预测未来，预测的准确性不太可能被检验。就像天气预报说有60%的几率下雨。如果没下雨，我们就不公平地说天气预报错了，当然我们并没有真正检验过概率解释的准确性。

相对频率视角。在这种与分析视角对立的视角框架下，事件的概率是通过收集数据，观察实际发生了什么及其发生的频率来计算的。如果我们将一对骰子掷上1000次，发现出现10、11或12点的几率是17%，那我们就会说得到这三个值其中一个的几率约是17%。

我们的陈述将是真正关于过去的解释，而不是对将来的预测。也许有人会说过去的事件能够对未来提供很好的参考，但是谁说得准呢？（那些对概率持分析视角的人，他们能够确定。）



HACK #4

1.4　否定虚无假设

实验科学家通过质疑向前推进。

科学是个目标驱动的过程，其目标是构建一个解释世界的知识体系。这个知识体系由一长串的科学法则、定律以及关于事物如何存在与运转的理论构成。实验科学引进新的法则和理论，并通过一系列逻辑步骤对其加以测试，这个测试过程称作*假设检验*。

1.4.1　假设检验

一个*假设*是对可检验世界做出的一个估计。比如，我也许会假设洗车导致下雨或是假设进浴

缸导致电话响了。在这些假设中，我认为洗车和下雨之间或洗澡和电话响铃之间存在关联。

验证这些假设是否正确的一个合理方式是观察假设中的变量（为了听起来更像统计学家，我们把这称作*收集数据*），看是否存在显在的相关性。如果数据显示变量间存在相关性，那么我的假设得到了支持，我也许有理由认为自己的估计是正确的。如果数据没有明显的相关性，我也许会明智地开始怀疑自己的假设是否错误，或是完全抛弃它。

科学家们通过收集数据来检验假设时，有4种可能的结果。表1-3显示了该决策制定过程中的可能结果。

表1-3：研究假设检验的可能结果

	假设正确：事实的确如此	假设错误：事实并非如此
数据支持假设：接受假设	A. 正确的决策：科学取得进步	B. 错误的决策：科学发展受阻
数据不支持假设：抛弃假设	C. 错误的决策：该死，又失败了	D. 正确的决策：科学取得进步

结果A和D可为科学的知识体系添砖加瓦。虽然A更可能让研究科学家激动不已，D其实也还不错。而B和C则是错误的，它们代表着错误的信息，只会混淆我们对世界的理解。

1.4.2　统计假设检验

你可能觉得假设检验的过程很有道理，这是一种相当直观的方式，可用来对世界和身处其中的人得出结论。人们在日常生活中总是通过这种假设检验来理解事情。

统计学家也检验假设，但针对的是某些非常特定的假设。首先，他们拥有代表样本值的数据，这些样本是从他们希望获得结论的真实或理论总体中抽取的。所以，他们的假设是关于总体的假设。其次，他们通常预先假设所关注总体内的不同变量之间存在某种相关性。统计学家提出的*研究假设*通常是这个样子的：所关注总体中变量X和变量Y之间存在相关性。

*统计假设检验*不同于*研究假设检验*，统计学家在假设检验结束时做出的概率解释，与研究假设为真的可能性无关。统计学家对研究假设为假的可能性做概率解释。在技术上更准确的表述为，统计学家对与研究假设相反的假设为真的可能性做出解释。这个相反的假设通常是关于变量间不存在相关性的假设，所以叫做*虚无假设*。统计学家提出的虚无假设通常是这个样子的：所关注总体中变量X和变量Y之间不具相关性。

研究假设和虚无假设涵盖了所有的可能性。变量间要么存在相关性要么不存在相关性。本质上，当必须从这两种假设中选择其一时，声明一个为假就等于为另一个提供了支持。因此从逻辑上讲，虚无假设检验和之前介绍的日常生活中人们自然运用的直觉方法一样有道理。研究人员执行虚无假设检验时偏好的结果和表1-3中介绍的一般假设检验方法略有不同。

如表1-4所示，统计学家通常希望否定他们的假设。统计研究人员通过否定虚无假设就能证实他们的研究假设，进而获得研究津贴，赢得诺贝尔奖，或许有朝一日他们的头像也能印在邮票上。

表1-4：虚无假设检验的可能结果

	虚无假设是正确的：总体中存在相关性	虚无假设是错误的：总体中不存在相关性
数据支持虚无假设：无法拒绝虚无假设	A. 正确的决策：科学取得进步	B. 错误的决策：科学进步受阻
数据不支持虚无假设：拒绝虚无假设	C. 错误的决策：该死，又失败了	D. 正确的决策：科学取得进步

尽管结果A在科学角度上是可以接受的，但在这种情境下结果D更令研究人员高兴，因为该结果支持了他们关于世界的真实推测，即研究假设。和上面一样，结果B和结果C仍然是有碍于科学进步的错误。

1.4.3 生效原理

统计学家检验虚无假设——估计希望找到的结果的反面，这么做有几个原因。首先，证明某件事为真的确非常非常难，尤其是当假设中包含了特定值时，这在统计学研究里是很常见的。证明一个特定的估计不正确要比证实它正确容易得多。例如，我无法证明我今年29岁，但要证明我今年不是29岁却相当容易。

同样，要证明某个特定的总体估计值不可能正确，也相对容易一些。大多数统计中的虚无假设都假定总体的值为0（也就是说，总体中变量*X*和变量*Y*之间不具相关性），要拒绝虚无假设就是去证明：不论总体的值是什么，它很可能不是0。对研究人员假设的支持通常来源于证明总体值大于零，并不需要具体说明这一总体值具体是多少。



对专业统计学家来说，这一点相当令人振奋，是不是？统计学家要做的就是告诉你，你的答案是错误的，无需提供正确答案！

甚至无需举出数字实例，科学哲学家长久以来便声称，科学是通过提出假设并试图证明假设错误而取得进步的。对于真正的科学来说，可证伪的假设是最好的假设。

统计分析通常按以下方式进行：提出与研究假设相反的虚无假设，然后尝试能否证伪该虚无假设。这个方法最早由20世纪早期最伟大的统计学家费歇尔博士（R. A Fisher）提出，此后便传播开来。此外还有一些其他的方法。许多现代统计学家认为我们应集中精力对所关注的总体值（比如变量间相关性的大小）做出最佳估计，而不是证明相关性大小为某个不确定的非零值。



1.5　增加样本量以减少误差

减少样本误差的最佳方式是增加样本量。

无论何时，只要统计学家使用样本而不是总体来做出推测，就必定会出现多多少少的误差。因为推断性统计的基本方法就是测量样本，并运用测量结果对总体进行估计[Hack #2]，我们知道这种对总体值的估计总会存在一些误差。好消息就是我们也知道如何把这些误差减少到最低程度。其解决方案就是增加样本量。

1713年雅各布·伯努利（Jakob Bernoulli）提出了一个适用于赌博情境的早期原则，称之为“黄金定律”。这条定律后来被其他人改称为“大数定律”（始于1837年法国数学家泊松）。它可能是统计学历史上最有用的发现，为所有研究者提供了关键的基本通用建议：*增加样本量！*



早期的应用统计科学（我们指的是17世纪和18世纪）几乎开口必提赌博和几率。这也许是因为它给那时代的绅士学者们提供了一个借口，打着智力追求的幌子行玩乐之实。当然，概率定律是统计程序和推论的数学基础，所以赌博应用很可能主要是用作统计概念教学中的最佳例子。

1.5.1　本定律的实际应用

本定律的一个应用是它对预估概率和实际发生概率的影响。它包含这样一个推论：对于受几率控制的结果，预测准确性的提高程度是一个固定的值。也就是说，预测准确性的提高程度是已知的。随着试验的次数增加，某个结果的预测概率和你观测到的实际发生概率之间的差距缩小，而且上述预期值和观测结果之间的差距大小可以计算出来。这种预期差距一般称为标准误差[Hack #18]。

结果的理论概率和实际发生概率之间的差距大小与以下的值成正比：

$$\frac{1}{\sqrt{\text{样本大小}}}$$

你可以把这则公式看作大数定律的数学表述。在概率和结果的语境下讨论准确性，样本量就是试验的次数。而在样本平均数和总体平均数的语境下讨论准确性，样本量就是样本中人的数量（或随机观测的数量）。

1.5.2　提高准确性

受此定律影响的特定值的大小取决于采用的测量尺度以及给定样本中变量的多少。然而我们可对样本量进行种种改变来提升推断的准确性。表1-5显示了所有推断性统计的准确性提升比例。依照定律可知：

表1-5：增加样本量的效应

样本量	误差的相对降低幅度	含　义
1	1	误差等于总体中变量的标准差
10	3.16	误差大小约为之前的1/3。观测的样本人数仅仅增加到10就极大提高了我们的准确性
30	5.48	样本人数从1 增加到30，会显著提高准确性。即便是从10增加到30都有作用
100	10	100个人组成的样本产生一个非常接近总体值（或者说期望概率）的估计。百人样本的误差大小仅为一个标准差的1/10
1000	31.62	根据如此大量的观测值产生的估计是相当准确的

1.5.3　生效原理

让我们从几个不同的角度来看这个重要的统计法则。我会用三种不同的方法来阐述这个定律，先从赌徒关心的角度开始，然后转到有关误差的话题，最后讨论采集代表性样本的意义。所有这些实际上说的是同一条规则，只不过阐述方式不同而已。

1. 赌博

如果一件事在单一试验中有特定的发生概率，那么它在无限次试验中的发生几率将与前述概率相等。随着试验次数接近无限，其发生概率将逐渐接近这个概率的值。

2. 误差

如果一个样本无限大，那么样本统计学特征就等于总体的参数。举例来说，随着样本量趋近无限大，样本平均数和总体平均数之间的差距逐渐缩小。随着观测数量的增加，总体值估计的误差逐渐变小，最终趋近于0。

3. 意义

相比从总体中抽取少数人的样本，抽取多数人的样本更具有代表性。随着样本量增加，样本所体现的总体的重要特征越来越多，同时预测的准确性也随之提高。



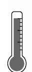
以上所有关于大数定律的论述成立的前提是：我们假设事件的发生或取样是**随机**的。

大数定律为标准误差的计算提供了基础，此外它还影响着其他一些核心统计问题，比如检验力[Hack #8]以及不该拒绝虚无假设时却加以拒绝的可能性[Hack #4]。雅各布·伯努利的赌徒伙伴们或许对他的黄金定律大感兴趣，因为这能让他们心中大致有数，知道还需要掷多少次骰子，掷出7点的可能性才会达到0.166或16.6%，然后据此制定较有把握的投注计划。

然而在过去的300年间，所有的社会科学无不利用这一简洁的工具，去估计用可见的事物来描述不可见事物所能达到的准确性。谢谢你，雅各布！

1.5.4　参阅

看清自己错到何种程度[Hack #18]



HACK #6

1.6　精确测量

测试需要综合诸多要素得出分数，经典的测试理论针对其中每一个要素提供了很好的分析。该理论的一个有用结果就是能对测试分数的精确程度进行估计和报告。

一个良好的教育或心理学测试产生的分数是有效度的和可信的。效度（validity）是测试分数能在多大程度上体现你希望测量的特征，以及对测量意图的有用程度。为了证明效度，你必须提供证据和理论来支持测试分数的解释是正确的。

信度（reliability）是对同一个人多次重复测量得到相同测试分数的一致性程度。要证明信度，就是要收集重复测量的数据并用统计学方法加以分析。

1.6.1　经典测试理论

经典测试理论，或者说信度理论，研究测试分数的概念。想想你某次参加测验的观测分数（你得到的分数）。按经典测试理论的定义，这个分数由两部分构成，用下列理论方程式表示：

观测分数=真分数+误差分数

这个方程式包含以下几个要素。

- 观测分数

你在测验中取得的实际分数。这通常等于正确回答的项目数，或者更通俗地说，测试中获得的点数。

- 真分数

真分数指你本该得到的分数。虽然这不是你应得的分数，或者说，这不是最有效的分数。真分数被定义为你无数次参加同一个测试，所得到的平均分数。注意，这个定义意味着真分数只是代表平均表现，或许能反映测试设计测量的特质，或许不能反映测试设计测量的特质。换句话说，一场测试也许会产生真分数，但未产生有效的分数。

- 误差分数

指你的观测分数和真分数的差距。

依据本理论，我们假定任何测试的分数表现都容易出现随机误差。你可能在自己实际上不知道答案的情况下，在社会学研究测验中猜对一道题。在这种情况下，是随机误差帮了你。



请注意，尽管这提高了你的分数，但依然是一个测量“误差”。

你也许在做早饭的时候遇到了一枚臭鸡蛋，结果心情不好，在应聘笔试时甚至根本没注意到最后那组题。这里，就是随机误差伤害了你。这些误差被认为是随机的，因为它们不是系统的，它们也和希望测量的特质无关。这些误差之所以被认为是误差是因为它们改变了你的分数，使之距真分数更远了。

如果进行多次测量，这些随机误差有时会提高你的分数，有时会降低你的分数，但是，纵观整个测试，误差的出现率应该比较平均。根据经典测试理论，信度[Hack #31]是测试分数随机波动的程度。代表信度的数字通常是通过观察测试中项目间的相关性来计算。这个指数范围分布在0.0和1.0之间，1.0表示一组没有任何随机误差的分数。指数越接近1.0，分数随机波动的程度越小。

1.6.2 标准误差的测量

尽管随机误差应该在多次测试情境下彼此消长达到平衡，但不完美信度依旧受到关注，因为决策几乎总是基于单次测试所得到的分数。比如在SAT考试中，如果你旁座的考生洒了古龙香水使你注意力不集中，结果考砸了，这种情况下，知道从长远来看自己的考试成绩会反映你的真分数也毫无意义。

测量学专家已经发明出了一套公式，用来计算你的真实分数水平落入的区间范围。这个公式利用了一个叫做“测量标准误差”的值。在一个测试分数总体中，测量的标准误差是每个人的观测分数与其真分数之间的平均距离。测量标准误差是利用测试的信度信息和群组观测分数的变异量（用标准差来反映）信息来估计的[Hack #2]。

计算测量标准误差的公式是：

$$\text{标准误差} = \text{标准差}\sqrt{1-\text{信度}}$$

以下的例子说明了如何应用这一公式。许多研究生院根据GRE测验的分数制定录取决策。GRE中的文字推理（Verbal Reasoning）的分数范围是200分到800分，平均值为500分（实际上，近年来的平均分比这个要低一点），标准差是100。

GRE测试分数的估测信度通常在0.92左右，这个值是相当高的。如果你参加GRE测验，得到了520分，那么恭喜你啦，高于平均分。520分是你的观测分数，但你的成绩容易受到随机误差影响。520分有多接近你的真分数呢？使用标准误差测量公式，可以计算如下：

(1) 1−0.92=0.08
(2) 0.08的平方根是0.28

(3) $100 \times 0.28=28$

GRE测验的标准误差约为28分，所以你的本次成绩520分很可能处于多次测验所得平均分上下28分的区间内。

1.6.3　建立置信区间

观测分数很可能在真分数的一个测量标准误差范围内，这是什么意思？如果有68%的几率，观测分数在真分数的一个测量标准差内，那么这是测量统计学家所接受的。然而应用统计学家喜欢超过68%，他们更愿描述为有95%的可能性包含真分数的观测分数区间。

想要说有95%的把握分数区间包含了个体的真分数，那么报告的分数区间应该是由加减大约两个测量标准误差构成。图表1-1显示了GRE的520的置信区间。

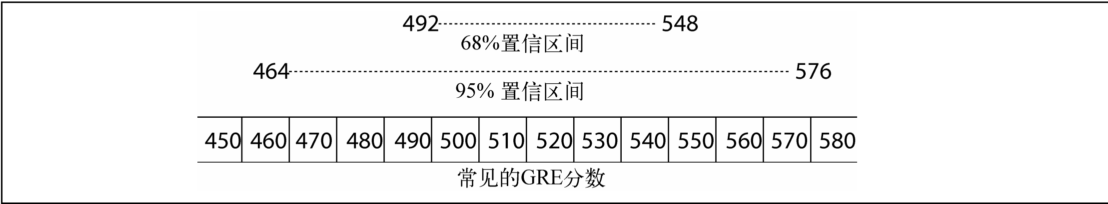


图1-1：GRE得分为520分的置信区间

1.6.4　生效原理

使用测量标准误差构建置信区间的方法是基于如下假设：误差（或误差分数）是随机的且这些随机误差呈正态分布。这里展示的正态曲线[Hack #25]就跟全世界凡有人类的地方所展示的一样。它的形状被大家所熟知并已被精确定义好了。有了精确性，就能够计算精确的置信区间。

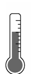
测量标准误差是一个标准差。在这种情况下，它是误差分数距真分数的标准差。在正态曲线下，68%的值都在平均数的一个标准差之内，95%的分数都大约在两个标准差内（更准确地说，是1.96个标准差）。就是这套广为人知的概率使得测量人员能够讨论95%或者68%的置信度。

1.6.5　意义讨论

知道测试分数95%的置信区间有什么用呢？如果是你要求学生参考并根据测试分数来做决策，那么你就能判断参考者的能力是否在你所设定的成功标准范围内。

如果你是参考者，那么你就能蛮有把握地知道自己的真分数在某个区间内。这可能会激励你再次参加考试，对自己凭运气可能取得的更好表现有一个合理的预期。如果你这次的GRE分数是520分，那么你就有95%的把握说，假如你马上再考一回，新分数可能会高达576分，当然，也有

可能低到464分。



HACK #7

1.7 提高测量尺度

四种测量尺度决定了利用测量所得分数的方式。如果你没有正确地应用测量尺度，那么就不能随心所欲玩转这些分数。

统计学方法分析数字。当然，这些数字必须有意义，不然的话，分析就没有多大价值。统计学者把有意义的数字称为*分数*。但是，统计学中使用的分数并不都“生而平等”。不同的分数因其生成时遵循的规律不同而载有不同的信息量。

当你决定测量某个对象时，必须谨慎选择赋值的规则。测量尺度决定了哪种统计分析是合适的，哪种是有效的，哪种是有意义的。



测量是对事物进行有意义的赋值。被测量的事物可以是具体的对象，比如岩石，也可以是抽象的概念，比如智力。

我们举个例子来解释“分数并不都生而平等”这句话的含义。假设你有5个孩子，都参加了一场拼写测试。满分是100分，查克得了90分，迪克和简都得了80分，鲍勃得了75分，顿只得了50分。如果有个朋友问你，孩子们在这场重要测验中表现如何，你可能会说他们平均分是75分。这是一个合理的概况总结。现在，想象一下你的5个孩子参加竞走对抗赛，这次是鲍勃第一，简第二，迪克第三，查克第四，顿第五。你那爱管闲事的朋友又问孩子们表现如何。你带着自豪的笑容说，他们平均拿到了第三名。这次就不再是合理的概括总结，因为它没提供任何信息。以上两种情况都使用了分数来表示成绩，其中的区别只在于选用的测量尺度不同。

一共有四种测量尺度，即四种以数字表示分数的方法，它们分别是*名义测量*、*次序测量*、*等距测量*和*等比测量*。不同尺度的区别在于所含信息量，以及在何种程度上可对其进行有意义的数学和统计分析。

1.7.1 将数字当做标签

如果你只打算用分数表示事物所属的不同类别，那就选择*名义测量*。名义测量仅把数字用做名称，即表示不同类别的标签。

比如，一名科学家收集了男女受试者的数据，他用数字1来表示男性测试对象，用数字2表示女性测试对象，这就是在名义尺度上使用数字。请注意，尽管数字2在数学上比数字1大，但在这个数据集里，2不代表更多，它只作为一个名称使用。

1.7.2　用数字来表示次序

如果你想用顺序或次序来分析你掌握的分数，那就选择*次序尺度*进行测量。次序测量提供了名义测量提供的所有信息，但增加了分数的次序信息。具有更大数值的数字能够和更小数值的数字比较，任何被测量的对象都能排出一个有意义的序列，不管是人或者海獭还是别的什么事物。

就拿你高中时的全班成绩排名作为例子：毕业典礼上致辞的最优生通常是平均成绩排第一的那个人。请注意，你可以和其他人比较名次，但是你不知道名次之间的具体分数差距是多少。在竞走比赛中，第一名可能只领先第二名1秒，而第二名可能比第三名领先30秒。

1.7.3　用数字来显示距离

*等距测量*涵盖了之前两个测量尺度的全部信息，并新增了准确性这一元素。这种测量尺度产生的分数，被认为在任何两个毗邻的分数间有相等的差距。

例如，在温度计上，70度和69度间的1度差距是有意义的，它完全等同于32度和31度之间的1度差距。这样的1度差距无论出现在温度计上的哪个位置，都代表着相同的热量值（你也可以说，温度计中液体所受的压力）。

等距尺度提供了比次序尺度更多的信息，现在你能对分数进行有意义的均隔了。大多数教育和心理测量都发生在等距测量这个尺度上。

虽然就我们在统计学上能做什么和不能做什么而言，等距测量看起来能解决所有的问题，但依然有一些数学运算在这个尺度上没有意义。比如，我们不用小数或比例进行比较。想想我们讨论温度的方式。如果昨天的气温是华氏80度，今天降到华氏40度，我们并不说“今天是昨天的一半热”。我们同样不说，一个IQ为120的学生比IQ为90的学生聪明1/3。



等距（interval）一词源自古代城堡建筑。你知道那些弓箭手防御驻守的角楼或塔楼吗？那些塔楼尖顶周围的圆形护墙，通常在每两块墙垛之间留一个射箭的垛口。这些垛口叫做等距（意为“两垛之间”），最好的防御设计就是等距设置墙垛和垛口，以提供360度的保护。

1.7.4　用数字来计数

*等比测量*作为测量的最高尺度，不但提供低尺度测量涵盖的所有信息，还能够进行比例比较和生成百分比。等比测量实际上是我们观察和考量自然世界最常用和最直观的方法。我们数数的时候，使用的就是等比测量尺度。邻居门厅里有几只狗？这个问题的答案也采用了等比尺度。

等比测量提供了如此之多的信息，而且能够进行所有可能的统计运算，因为比例尺度拥有绝对意义的零点。这个绝对意义的零点意味着在刻度尺上一个人可以得0分，在被测量的特征上真正得0分。虽然温度计上也有一个0，但气温为0度不意味着绝对没有热量。在等距刻度上，比如我们的温度计上，分数可以是负值。但在等比测量尺度中却没有负值。

1.7.5 选择合适的测量尺度

哪个测量尺度适合你？因为达到等距尺度所得到的优势，所以大多数社会科学家倾向于在等距和等比尺度上进行测量。在等距尺度，你能够安全地进行描述性统计，执行推断统计分析，比如t检验、方差分析以及相关性分析。表1-6概括表现了每个测量尺度的优点和缺点。

表1-6：测量尺度

测量尺度	优　　点	缺　　点
名义测量	描述分类数据	数据不代表数量
次序测量	允许分数间比较	很难概括分数
等距测量	可进行大多数统计分析	不能进行比例比较
等比测量	绝对零点使得所有的统计分析都能完成	有些变量没有绝对零点

为了对他人的研究数据选取正确的统计分析方法，需要先识别他所使用的测量尺度并利用该测量尺度的优势。如果是你自己生成数据，可以考虑提高测量尺度：采用尽可能高的测量尺度。

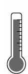
1.7.6 具有争议的工具

自从20世纪50年代测量尺度概念被广泛接受以来，一直存在一些争议，即我们是否真的有必要等距尺度上执行统计分析。有很多常用的测量形式（比如态度量表、知识测验或人格测验）不是确定无疑地在等距尺度上进行的，它们也许处于次序测量的最顶端。我们能在分析中安全地使用这些本该在等距尺度上获得的数据吗？

研究文献中的一个主要共识是：如果你至少处于次序尺度，而且有把握能对等距尺度统计分析做出解释，那么你就可以在这种类型的数据上安全地执行推断性统计分析。顺带说一句，在现实研究中，几乎每个人都在有意识或无意识地采用这个方法。

但是，我们很难否认依据测量尺度进行统计分析决策这一基本理念。一个足以说明测量尺度重要性的经典例子是费雷德里克·洛德（Frederick Lord）在1953年发表的论文“橄榄球运动数字处理的统计”（“On the Statistical Treatment of Football Numbers”,《美国心理学家》杂志，Vol. 8, 750~751）。这位粗心的统计学家热切地分析了一些所关注球队的数据，写出了一篇满是平均数和标准差以及其他一些复杂分析的报告。但后来发现，这些数据竟然是运动员所穿运动衫上的数字。

也许，这是一个没有注意到测量尺度的明显例子。但这个统计学家仍然力挺自己的报告。“这些数据本身不知道自己从哪儿来，”他辩解说，“但它们依然有效。”



HACK #8 1.8 提高检验力

在社会科学研究中，成功通常被定义为发现了统计显著性。为提高做出任何发现的几率，有统计见识的超一流科学家应当主要致力于提高检验力。

进行基于统计的研究，会遇到两种潜在的陷阱。科学家们可能认定自己在总体中发现了什么，但这种东西其实只存在于他们手头的样本中。反之，他们也可能在样本中什么都没找到，但实际上，总体中确实存在极妙的相关性，只待他们去发现。

第一个问题可通过代表取样而最小化[Hack #19]。第二个问题可通过提升统计检验力来解决。

1.8.1 检验力

在社会科学研究中，统计分析总要判断样本中观测到的某个值有没有可能是随机发生的。这个过程称作显著性检测。显著性检测产生一个*p*值（概率值），表示样本可以从特定的相关总体中抽取的概率。

*p*值越低，我们就越有信心认定，相关结果具有统计显著性，而且数据揭示出的相关性不仅存在于样本中，也存在于其代表的总体中。通常来讲，会对测量的事物选择一个预先设定好的显著性水平作为标准。如果最后*p*值等于或小于预先设定的显著性水平，就表明研究达到了一定的显著性水平。



统计分析和显著性检验并不局限于确认变量间的关系，借助一些最常见的分析（t检验、F检验、卡方检验、相关系数、回归方程等）通常能达到这一目的。我在此讨论相关性，是因为这是你所期望的典型效应。

统计检验的效力是指：假定总体内的变量间存在相关性，统计分析达到显著性的概率。注意这是一个条件概率。总体中必须存在相关性，否则，检验力就毫无意义。

检验力不是找到显著性结果的几率，它是在相关性存在的前提下，找到相关性的几率。检验力公式包含三个组成部分：

- 样本量；
- 预设要达到的（需小于）显著性水平（*p*值）；
- 效应值（总体中相关性的大小）。

1.8.2 执行检验力分析

假设我们要对比两组不同的样本，看它们之间是否存在足够的差异，能够说明二者各自代表的总体间确实存在差异。比如，假设你想要知道男性和女性谁的睡眠时间长。

这个设计非常简单。创造两组样本群：一组男性，一组女性。然后，调查两组人，问他们每晚通常睡几个小时。但是，为了找出真正的差异，你需要调查多少人？这就是一个检验力的问题。



t检验比较两组样本分数的平均表现，看是否存在显著差异[Hack #17]。在这种情况下，统计显著性意味着这两组样本所代表的两个总体间的分数差异很可能大于零。

在研究开始前，研究人员可以决定统计分析中使用的检验力。为了计算检验力，需要知道三样东西，但其中两样在研究开始前就已经知道了：你能决定样本量以及选择预设的显著性水平。你所不知道的是变量间相关性的实际大小，因为计划中的研究结果数据还没有产生。

研究人员能在研究开始前对所关注变量之间的相关性大小（即效应值）进行估计，检验力同样可以在研究开始前被估计。通常来讲，研究人员会对最重要或最感兴趣的方面设定最小相关水平。

一旦这三样（样本量、显著性水平和效应力）都确定了，便可计算第四样（效应力）了。实际上，在这四样东西中，设定了任何三样的水平，都能计算出第四样。比如，一名研究人员通常知道分析中需要的检验力大小、报告具有统计显著性所需的效应值、选择的预设显著性水平。有了这些信息，研究人员就可以计算出需要的样本量。



为了估计检验力，研究人员经常使用一个得到普遍接受的标准方法，其中将检验力目标值设为0.80，将预设显著水平设为0.05。检验力水平在0.80，意味着总体中**如果存在**相关，那么研究人员会有80%的几率在样本中发现相关性或效应。

t检验中，效应值（或者相关性大小指数[Hack #10]）常用两组平均数差除以样本标准差所得的值来表达。如此得出的效应值，0.2以内视为小，0.2~0.5视为中，0.8视为大。效应力分析需要解决的问题是：这两组中各需要多大样本（多少人）才能在测试分数中找到显著性差异？

计算检验力的推导方法很复杂，在此就不予介绍了。在现实生活中，我们估计检验力一般是运用计算机软件，或者查考统计书后所附的密密麻麻的表格。不过，我算出了各种选项的效应值，呈现在了在表1-7里。注意关键变量是效应值和样本量。依据传统习惯，我把检验力设置为0.80，显著水平设为0.05。

表1-7：不同效应值所需样本量

效应值	样本量
0.10	1600
0.20	400
0.30	175
0.40.	100
0.50	65
1.0	20

想象一下，在你的“性别与睡眠”研究中存在实际差异，但很小。t检验分析中将大约0.2标准差的组间差异视为小差异，所以你可能会预期效应值为0.2。为了发现这个小的效应值，每组的样本量需要达到400人！随着效应值的增加，所需样本量变小。如果总体效应值是1.0（一个非常大的效应值，两组间存在巨大差异），每组20人就足够了。

1.8.3　推测极妙的相关性

科学家总是依赖统计推论来拒绝或接受他们的研究假设。他们总使用虚无假设，先设定变量间不具相关性或组间没有差异性。如果样本显示总体中的变量间实际上存在相关性，他们就会拒绝虚无假设[Hack #4]，接受备择假设，即他们的研究假设，作为对现实的最好估计。

当然，这个过程中可能出现错误。表1-8列出了在假设检验游戏中可能出现的错误类型。当你不应该拒绝虚无假设时你却拒绝了，统计哲学家们将这种错误称为I型错误。当你应该拒绝虚无假设的时候，却没有拒绝，这被称为II型错误。

表1-8：假设检验中的错误

行　　为	虚无假设是对的	虚无假设是错的
拒绝虚无假设	I型错误	显著性发现
接受虚无假设	正确的决策	II型错误

作为一名聪明的科学家，你要做的是避免这两类错误，并发现显著性。当虚无假设是正确的，接受虚无假设，获得正确的决策也不错，但这没有发现显著性那么有趣。“把你的一生贡献给表格中的右上象限吧，”我叔叔弗兰克经常说，“你将变得超乎想象地开心和富有！”

要想加大发现统计显著性的几率，一个在你控制之外的条件必须为真；那就是，虚无假设必须为假，否则“发现”什么的几率就少得可怜。此外，如果你“发现”了什么，但它实际上并不存在，你就犯下了严重的I型错误。在总体的研究变量间必须确实存在相关性，这是你在样本中发现这种相关性的前提。

所以，你最终是否落在表1-8右列中，完全取决于命运。检验力是一旦你到达右列就移到顶格的几率。换句话说，检验力是当虚无假设为谬时，正确拒绝虚无假设的几率。

1.8.4 生效原理

效应值和样本量之间的关系是有意义的。想象有一种动物躲在干草堆里（这动物是效应值，拜托，只在我的这个比喻中有效）。你只需较少的观察（撩开几把干草）便可发现大的效应值（比如一头大象），这要比发现一种小动物（比如像可爱的水獭幼崽）方便得多。人数代表观察数，隐藏在总体中的大效应值比小效应值更容易发现。

检验力中效应值和样本量的普遍关系，反过来也同样有效。在已知的效应值下估计，只消提高样本量，到一定程度就会拥有你所需的检验力。记住，表1-7假定你想要80%的检验力。你可以采取较小的样本量，只是会有较低的检验力。

1.8.5 不适用领域

记住检验力不等于成功的几率，这很重要。它甚至不是达到某个显著水平的几率。它是在研究者的所有估计值都是正确的情况下，达到某个显著水平的几率。这公式最难估计或设置的部分是总体中的效应值。研究者很少知道自己在找寻的事物相关性有多大。归根到底，如果他知道研究变量间相关性的大小，那就没有做研究的必要了，是吧？



1.9 展示因果

统计研究人员已经建立了一些基本原则，如果你希望证明一件事情是另一件事情的原因，那你就得遵守这些原则。

使用统计数据的社会科学研究有着广泛的目标。其中一个目标是收集和分析有关世界的数据，用来支持或否定变量间关系的假设。第二个目标是检验假设，看变量间是否存在因果相关。与目标二相比，目标一是件容易的事。

世间万物之间存在各种各样的关系，统计学家也发明了各种方法来找到这些关系，但是相关性的存在并不意味着某个特定变量是另外一个变量的原因。比如，人群中身高和体重之间存在良好的正相关[Hack #11]，但是如果我瘦几磅，我不会变矮。反过来说，如果我长高了几英寸，我的体重很可能会增加。

只知道两者相关，并不能真正告诉我一件事是否导致另一件事的发生。不过，相关性缺失似乎能说明因果方面的问题。如果两个变量间不存在相关性，似乎就能排除一个变量是另外一个变量原因的可能性。相关性存在使得因果关系有可能存在，但无法证明它的存在。

1.9.1　设计有效的实验

研究人员已经发展出一种框架，用来讨论各种研究设计，以及这些设计是否有可能证明一个变量对另一变量产生影响。不同的设计在于有无对照组以及被试如何分配。

基于设计能否提供因果关系的强证据、中等证据、弱证据或是无证据，共有四种基本的组设计类别。

- 非实验设计

这类设计通常只包含一组人，统计数据常被用来描述总体或是证明变量间关系。这种设计的一个例子是相关性研究，分析变量间简单的关联[Hack #11]。这种类型的设计并不提供因果关系证明。

- 预实验设计

这类设计通常对一组人运用两套或更多的测量手段，看结果是否有所改变。这个设计的一个例子就是对一组人进行预测试，对他们做点什么，然后对他们做一次实验后测试，看他们的分数是否发生了改变。这种类型的设计提供了很弱的因果关系证明，因为除了你对这些人施加的作用外，还有其他的外力可能会导致分数的改变。

- 类实验设计

这种设计包含不止一组人，至少会有一组作为对照组。对各组成员的分配不是随机的，而是通过研究者无法控制的一些东西决定的。这种设计的一个例子就是对比男性和女性对统计学的态度差异。最好的情况下，这种设计能提供因果关系中等强度的证明。如果没有随机分组，各群组很可能在一批未测量的变量上不等同，这种不等同可能是导致所发现差异的真正原因。

- 实验设计

这类设计有一个对照组，重要的是，被试是随机分配到各组的。随机分配被试使得研究人员可以假定所有的群组在未测量变量上是等同的，因此（在理论上），如果发现任何差异性，则把它们剔除作为备择解释。这种设计的一个例子就是药物研究，所有被试被随机分配到两组，一组服用药物，另一组作为对照组，服用安慰剂（糖丸）。

1.9.2　体重会影响身高吗

在本条Hack的稍前部分，我提到关于相关性的一个著名发现：在人群中，身高和体重似乎存在相关。比如，个子高的男性胖些，个子矮的男性瘦些。我觉得这个说法挺滑稽的，因为如果真是这样，那么只要给他们多吃点，他们就会长高一些。因为我知道身体发育的原理，所以体重会影响身高在理论上不太可能。但如果你要科学的证据，那该怎么证明呢？

我可以通过一个基本的实验设计，来检验“体重会影响身高”这一假设。实验设计必须有一个对照组，且被试的分配必须随机。在这种情况下发现的任何相关都可能是因果相关。在我的研究中，我会创建两个组。

- 第1组

30名大一新生，是从我工作的中西部大学（Midwestern University）这一总体中招募的。这组是实验组，我会增加他们的体重，然后测量他们的身高是否有所增加。

- 第2组

30名大一新生，也是从我工作的中西部大学这一总体中招募的。这组是控制组，我不会对他们的体重进行任何操控，然后测量他们的身高是否发生改变。



在这一设计中，科学家将体重称作**独立变量**（因为我们不关心是什么引发体重变化），将身高称作**因变量**（因为我们想知道它是否**依赖于**独立变量，或是由独立变量引发的）。

因为这一设计符合实验设计的标准，所以我们能将所发现的任何相关视为因果关系的证明。

1.9.3 抵御对效度的威胁

研究结论分为两类。它们关系到可否做出因果声明，以及该因果声明确立后，可否被推广到整个总体或是实验室之外。表1-9展示了解释研究结果时需要考虑的效度类型。研究人员就好比跨栏运动员，必须跨越这些栏杆。

表1-9：研究结果效度

效度考虑	效度问题
统计结论效度	变量间存在相关吗
内部效度	该相关是因果相关吗
构建效度	变量间的因果关系是否因此而受到影响
外部效度	这因果关系到处都存在吗

即便研究人员选择了真正的实验设计，他们依然要担心结果实际上或许不是由于一个变量对另外一个的影响造成的。对因果结论的效度造成威胁的因素有很多，但幸好，研究人员只需要想想，就辨识出很多这样的威胁并开发了解决方案。



研究人员对群组设计的理解、用来描述它们的术语、研究设计中对效度威胁的确认以及应对威胁的方法，几乎完全来自于Cook和Campbell两人影响深远的论著，见1.9.4节。

以下部分将讨论针对因果声明以及普遍性声明的威胁，并讲到若干消除威胁的方法。在已有研究文献中识别出的威胁有几十种，也试图给出应对方法，但其中大多数要么无法解决，要么可以用这里介绍的工具来解决。

- 历史

外界事件能够影响结果。一个解决办法是使用控制组（一个未接受药物或任何干预的对照组），并将被试随机分配。这个方案的另一部分是尽可能控制两组的环境（比如在实验室环境下）。

- 生物成熟

在一项研究过程中，被试自然地生长发育，所以改变有可能是这种自然发展导致的。对实验组和控制组随机分配被试能很好地解决这个问题。

- 选择

在分配被试时可能存在系统性偏差。解决的办法是随机分配被试。

- 测试

只进行一场预测试也许会影响研究变量的水平。创建一个对照组，对两组都进行预测试，这样两组中的任何改变都是等同的。还有，要对两组随机分配被试。（你开始看出某种模式了吗？）

- 测试设备

在测量中可能会有系统性偏差。解决办法是使用有效的、标准化的、客观的分数测试。

- 霍桑（Hawthorne）效应

被试知道自己在参加实验的意识也许会影响结果。为了应对这个问题，你可以限制被试对你期望的实验结果的认知，或者执行一场双盲研究（double-blind study），即被试和研究人员都不知道给予被试的是什么刺激。

研究设计的效度以及任何一种因果关系声明的效度，都类似于测量中的效度声明[Hack #28]。这样的讨论是开放和无止境的，效度结论依赖于对手边证据的合理检验，以及对什么结论看似合理的考虑。

1.9.4　参阅

- Campbell, D.T. and Stanley, J.C. (1966). *Experimental and quasi-experimental designs for research*. Chicago: Rand McNally.
- Cook, T.D. and Campbell, D.T. (1979). *Quasi-experimentation: Design and analysis issues for field settings*. Boston: Houghton-Mifflin.

❑ Shadish, W.R., Cook, T.D., and Campbell, D.T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston: Houghton-Mifflin.



1.10 敏锐识别效应值

你刚读到一条令人震惊的科学新发现，但这真的是一个重大发现吗？通过对效应值的解读，你能判断这类声明（或缺乏这类声明）对你究竟有多重要。

大多数非科学出版物、电视上、电台里，甚至网络上报道的科学发现总是缺少了点什么。虽然这些媒体都很擅长报告“统计显著性”，但这并不足以判断是否存在真正重要的或是有用的发现。一项大规模药物研究可以报告“显著”结果，但仍未发现任何令我们或其他研究者感兴趣的东西。

正如本书中一再指出的那样，显著性[Hack #4]只是意味着，你关于样本的发现在总体中可能为真。问题是，仅有这个事实并不足以让你知道自己是否应当改变行为，开始一种新的饮食方式，改变药物或者重新解读你的世界观。

要想根据任何新的科学报告对生活和现实做决策，你需要知道刚被揭露出的相关性有多大。品牌A比品牌B究竟要好多少？用有意义的话来表达，男孩和女孩的SAT成绩差异到底有多大？每天服用半片阿司匹林用以降低心脏病发作的风险，这么做值得吗？即便的确能降低上述风险，那么究竟能降低多少？

这种相关性的强度也应该以某种标准化方式来表达，否则就没有办法切实判断它的大小。使用效应值这一统计工具，能让你敏锐识别效应的大小。

1.10.1 效应值无处不在

效应值是一个标准的值，表示两个变量间相关性的强度。在讨论如何辨别或是解读效应值之前，让我们先学习一些关于相关性和统计研究的基础知识。

统计研究总是对变量间的相关性感兴趣。比如相关系数，它是体现两组分数间关系强度和方向的指数[Hack #11]。测量关系的统计方法包括t检验[Hack #17]和方差分析，这是一次性对比多组的一种方法。它们虽然不是很明显，却依然有效。



即使是用来比照不同组的方法，依旧对变量间的关系感兴趣。比如，在t检验中，一个显著性结果意味着一个人在哪个组至关重要。换句话说，在独立变量（定义群组）和因变量（测量结果）间存在关联。

1.10.2　发现或计算效应值

这个Hack关于发现和解读效应值，以便判断大众媒体或科学论著所报告的科学发现的意义。通常情况下效应值会直接报告出来，你只需要知道如何解释它。其他时候，虽没有报告效应值，但提供了足够的信息能让你算出效应值。

效应值的报告方式通常有三种类型。它们的区别在于使用方法的不同以及这些方法量化信息的方式不同。在每种情况下，效应值可被解读为对“变量间关系的大小”的估计。下面分别对三种类型的效应值加以介绍。

- 相关系数

相关，用小写r来表示，其本身已经是对变量间关系的度量，所以它是一种效应值。因为相关可以是负的，所以，有时候会对r值进行平方得到一个大于0的值。因此，r^2被解释为变量共享的“方差比例”。

- d

这个值用d来表示，够奇怪的，它归纳了t检验中所使用的两组平均数的差异。其计算是通过两组平均数的差异除以两组平均标准差而得到的。



这儿还有另一种计算d值的方式，简单、超级有趣、相当酷、干净利落：

$$d = t\sqrt{\frac{\text{第1组样本大小}+\text{第2组样本大小}}{(\text{第1组样本大小})(\text{第2组样本大小})}}$$

- η^2

方差分析结果中报告的效应值最常用η^2来表示。跟r^2类似，它被解释为因变量（结果变量）对独立变量（你所在的组）贡献的“方差比例”。

1.10.3　解读效应值

关于显著性水平，统计学家们已经有了一些判别何为“良好”的指标。比如，大多数统计研究人员希望显著性水平达到0.05或更低。尽管在效应值的问题上，并不总是存在明确的好坏之分，但要分辨效应值是大、中还是小，依然有一些标准可循。

绝大部分情况下，判定大、中和小的标准，应视实际研究中通常发现的效应值而定。如果某个特定效应值大到极少见于已有的研究结果，那它就被视为大。如果效应小且在实际研究中常见，那么它就被视为小。

然而，在解读研究结果时，你应该自己决定，效应值达到多大才能引起你的兴趣。这取决于调查研究的领域。表1-10提供了判别效应值的经验法则。

表1-10：判别效应值的标准

效应值	小	中	大
r	+/−0.10	+/−0.30	+/−0.50
r^2	0.01	0.09	0.25
d	0.2	0.5	0.8
η^2	0.01	0.06	0.14

1.10.4 解读研究发现

在讨论研究结果时，关注效应值的好处在于：能让每个人都大致了解，给定的研究变量（或干预、药物、教学技术等）对现实的实际影响有多大。因为报告效应值的时候，通常不会报告概率信息（显著性水平），所以搭配了显著性水平报告的效应值是非常有用的。这样，你可以回答两个问题。

- 这种相关可能存在于总体中吗?
- 这种相关有多大?

还记得前面的例子吗？你应该每天服用半片阿司匹林以降低心脏病发作的风险吗？20世纪80年代末有一项广为人知的研究，发现了这两个变量在统计学上的显著相关。当然，在做任何类似的决定前，你应该先和你的医生谈谈，但你同样应该获取尽可能多的信息来帮你做出决定。现在让我们借助效应值信息来解读此类发现。

以下是媒体的报道：

> 22 071名内科医生组成的样本被随机分为两组。很长一段时间内，半数医生每天服用阿司匹林，同时另外半数服用外观和味道与阿司匹林相似的安慰剂。实验期结束时（实际上较早就结束了，因为大家认为阿司匹林的药效实在很强），服用阿司匹林的医生患心脏病的几率为服用安慰剂组医生的一半。服用安慰剂组，有1.71%的人有心脏病发作，而阿司匹林组医生的发病比例仅有大约1%（0.94%）。这一发现在统计学上具有显著性。

对上述发现的"清楚"解释是，服用阿司匹林可以使心脏病发作的几率减半。假设这个研究具有代表性，而且参与其中的医生们和你我极其相似，那么这个解释可谓相当正确。

对于该发现的另一种解读方式，是看阿司匹林服用的效应值。通过比例比较公式得到，该研究的效应值为0.06个标准差，即d为0.06。根据表1-10中的效应值判别标准，该效应值应该解读为小，真的非常小。这样的解读表明在服用阿司匹林和心脏病发作之间的确存在一个非常小的相关。相关性确实存在，只是不太强。

这个问题还可以这样看：首先，在一段给定期间内，心脏病发作的几率相当小。研究中98.76%的人没有心脏病发作，不管他们是否服用阿司匹林。虽然服用阿司匹林的确会降低心脏病发作的几率，但这个下降只是从微小到更小一点而已。一个类似的情形是：和没有购买彩票的人相比，如果你大量购买彩票，中奖几率会进一步提升，但这个几率依然很小。

1.10.5　生效原理

一个研究人员可能取得显著性结果，但仍然没有发现任何令人激动的东西。这是因为显著性只能告诉你样本结果很可能不是偶然发生的。这结果是真实的，也可能存在于总体中。如果你在两个变量间或是用药和治疗结果间找到细微的相关，那么这种相关可能由于太小以至于没人对其真正感兴趣。药效或许是真的，但很弱，所以不值得推荐给病人。A和B的相关可能大于0，但它的值还是太小，对理解两个变量中的任何一个都没有太大的帮助。

现代研究人员依然热衷于寻找自己的发现中是否存在统计显著性，但他们应该几乎总是报告并讨论效应值。如果报告了效应值，你就能够解读它。如果没有报告效应值，你也总是能从公诸于众的研究报告中挖掘所需信息，自己计算出效应值。其中的绝妙之处在于，你也许比报告这些发现的媒体，甚至比搞这项研究的科学家自己，更懂得该发现的重要性。

第 2 章

发现相关性

（*Hack #11～#22*）

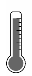
我们周围存在着无形的关系网。变量A引发变量B，变量B影响了变量C，变量C完全独立于变量D，除非变量E也参与进来。本章介绍的Hack能让你发现这些联系并准确描述它们。这些Hack揭示了人们做事的内在原因和事物之所以成为现在这个样子的原因。

一个特质与另一个特质之间的联系，因与果之间的联系，都是可以通过正确的技巧轻松揭示的关系。我们从确认任何关联的强度[Hack #11]开始，然后画出它的样子[Hack #12]。接下来，用你所掌握的相关性知识进行预测[Hack #13]，再提高这些预测的准确性[Hack #14]。有些相关是通过观察非预期结果的发生[Hack #15和Hack #16]或注意组间的真正差异[Hack #17]而显现出来的。

因为我们无法测量自己可能感兴趣的每个人、每条鱼或是每棵松树，所以要依赖有代表性的样本[Hack #19]为我们提供观测值。然而，样本有可能产生误导[Hack #18]，也可能以令人意想不到的绝妙方式起作用[Hack #20]。

要想与别人分享你的发现或理解这些发现对你的意义，你需要注意避免受骗，也不要欺骗他人。小心不要误解任何数字[Hack #21]或图像[Hack #22]。

将这些方法统统打包带好，把自己武装起来，去发现那些有待发现的事物吧。



HACK #11

2.1　发现相关

揭示世间各种无形的联系，不过是记录观测值并计算出那些奇妙而神秘的相关系数而已。

关于人们为何产生如此这般的感受、做出如此这般的事情，你可能会做出形形色色的假设。统计研究人员把这些假设称作变量间相关性的假设。

不管科学界怎么称呼它，你在现实生活中很可能就是这么做的。你可能会对态度和行为、态度和态度或是行为和行为之间的关联进行估计。你可能试图理解周围世界中的人，因而随便做出假设；或者你是一名市场营销专家，需要借此来理解顾客；又或者你是一名心理学研究生，正为完成一项针对自尊和抑郁的相关性进行统计分析的课堂作业而伤脑筋。

在统计学里，这样的关系称作*相关*。描述关系大小的数字是*相关系数*。通过计算这个有用的值，你能够获得任何有关“关系”问题的答案（除了恋爱关系，那只能靠你自己了）。

2.1.1　检验关系假设

想象有这么一个研究：美国奶酪蛋糕零售协会（American Cheesecake Sellers Association）的一名研究人员做出假设，认为人们喜欢奶酪蛋糕是因为爱吃奶酪。也就是说，他猜测人们对奶酪的态度和对奶酪蛋糕的态度之间存在相关。如果他的假设最终被证明是对的，那么他将从美国奶酪爱好者协会（American Cheese Lovers Association）购买大量的邮寄地址，向这些人发送宣传册，介绍奶酪蛋糕的保健功效。如果他是对的，销售量将如火箭般蹿升！

为了检验自己的假设，他创建了两项调查研究。其中一项是让受访者表述对奶酪的感觉，另一项则询问他们对奶酪蛋糕的感觉。50分表示这个人喜欢奶酪（或奶酪蛋糕），0分表示这个人讨厌奶酪蛋糕（或奶酪）。表2-1显示了他上班途中在公交车上收集的5个人的数据。

表2-1：关于对待奶酪和奶酪蛋糕的态度之间相关性的数据

受访者	对奶酪的态度	对奶酪蛋糕的态度
拉里	50	36
莫伊	45	35
乔	30	22
塞夫	30	25
格劳乔	10	20

看看这些数据，两个变量间看起来是否存在相关？（看吧，我会给你30秒时间。）

我会说二者之间存在一种非常清晰的关系。在奶酪量表上得分高的人，同样在奶酪蛋糕量表上得分高。当然，这些人在两个量表上的得分并非完全相同，甚至分数高低顺序也不相同，但是相对来说，每个人在两张态度量表上相对于其他人所处的位置大致相同。那位奶酪蛋糕零售协会的研究人员为他的假设找到了支持。

2.1.2　计算相关系数

只对样本中的两列数据扫上几眼，并不足以确知两件事之间是否存在相关。在这个例子中，

市场营销专家想用一个数字更加准确地描述所发现的关系。

相关系数考虑了我们在观察表2-1中两列数字时使用的所有信息，并判断其间是否存在相关。相关系数的计算公式包括以下几个步骤。

(1) 查看一列中的每个分数。
(2) 查看每个分数和本列平均数的距离。
(3) 查看另一列中与其对应的分数与平均值的距离。
(4) 将这一对距离数字相乘。
(5) 计算乘积结果的平均数。

如果这是本统计教科书，我就有必要展示略为复杂的相关系数计算公式。称它“略为复杂的”算是轻描淡写。坦白讲，那些公式非常可怕。相信我，为了你好，我不会把这些可怕的公式展示给你看，而是展示一个看起来令人愉快的、友好的公式（而且同样有效）：

$$\frac{\sum\left(Z_x Z_y\right)}{N-1}$$

其中Z表示Z分数，是一个分数离平均数的距离。随后，将这些距离除以分布的标准差。因此，Z_x表示第一列的所有Z分数，Z_y表示第二列的所有Z分数。Z_xZ_y表示将它们相乘。Σ符号表示相加。所以，此方程的意思是把所有配对的Z分数相乘，并把这些乘积相加，然后除以配对数（N）减1。

平均数是一组分数的算术平均。其计算方法是将所有数字相加并除以分数的总数。一组数的标准差是各个分数距平均数的平均距离。

在使用我们的相关公式计算Z分数前，我需要知道每列数据的平均数和标准差。计算这些关键值的公式在“仅用两个数字描述世界”[Hack #2]中已有介绍。以下是本例中两个变量的平均数和标准差。

- 对奶酪的态度

平均数=33；标准差=15.65

- 对奶酪蛋糕的态度

平均数=27.6；标准差=7.44

表2-2给出了一些针对奶酪态度数据所做的计算。

表2-2：对于奶酪和奶酪蛋糕的态度之间相关性的计算

受访者	对奶酪的态度	对奶酪蛋糕的态度	奶酪的Z分数	奶酪蛋糕的Z分数	Z分数的乘积
拉里	50	36	1.09	1.13	1.23
莫伊	45	35	0.77	0.99	0.76
乔	30	22	−0.19	−0.75	0.14
塞夫	30	25	−0.19	−0.35	0.07
格劳乔	10	20	−1.47	−1.02	1.50

相关系数为0.93，非常接近于1。1是最强的正相关，所以人们对奶酪–奶酪蛋糕的态度之间存在非常强的相关。

2.1.3　解释相关系数

有点神奇的是，相关公式的计算产生一个范围在−1.00~+1.00的数，用以表明两个变量间的关系强度。正号（+）表示正向相关，即随着其中一个值的增加，另外一个值也增加。负号（–）表示反向相关，即随着其中一个值的增加，另外一个值减少。需要指出的一个重点是：相关系数提供的是两个变量间线性关系强度的标准度量[Hack #12]。

相关的方向（不管是正的还是负的）是标尺方向的虚拟结果，人们选取这个标尺来度量变量。换句话说，强相关也可以是负的。就拿高尔夫球技和高尔夫平均得分之间相关性的度量来说，球技越高，分数越低，但你依然可以预见二者之间存在一个强相关。

2.1.4　统计显著性和相关

我们的市场营销专家可能同样对“样本的相关性是否大到有可能抽取自相关性大于零的总体”这个问题感兴趣。换句话说，我们在样本中发现的相关是否足够大，以至于它肯定来自于一个变量间至少存在某种关系的总体?

相比从小样本（比如前述的5位公交乘客）中得到的相关，本例中的市场研究人员更相信由大样本中获得的相关。如果他将这个相关呈报给老板，但结论对大多数人并不适用，那么他说不定要被炒鱿鱼，只能靠开小型客货车卖奶酪蛋糕来谋生了。

表2-3展示了样本中相关系数必须达到多大，统计学家才能够确定其代表的总体中存在大于零的相关。

表2-3：可能并非偶然出现的相关

样本量	可被视为统计相关的最小相关
5	0.88
10	0.63

（续）

样本量	可被视为统计相关的最小相关
15	0.51
20	0.44
25	0.40
30	0.38
60	0.26
100	0.20

就我们的5人样本而言，任何大于或等于0.88的相关系数会被认为统计学显著（意思是“相关性大到很可能存在于样本所代表的总体中”）。

2.1.5 其他生效领域

只要满足某些条件，你就能够计算作为两个变量间关系强度度量的相关系数。

- 你一定能够以这样一种方式测量变量：数字有实际的意义且能够代表一些基本的连续概念。连续变量的例子有态度、情感、知识、技能，那些你能够计数的事物，比如因为喜爱奶酪蛋糕导致体重增加的磅数。（如果你测量的事物不是连续的，就像存在不同类别的情况，比如性别或党派，你依然能够计算相关，只不过不用这里展示的公式。）
- 变量必须真正存在差异。如果每个人对奶酪的感觉都一样，你就不能计算对奶酪（也可以是巧克力或其他什么）的态度相关。数学需要差异性。
- 只有在样本是从总体中随机抽取的情况下，达到统计显著所需的最小相关系数大小（如表2-3所示）才是精确的。研究人员（比如我们的奶酪蛋糕营销人员）必须判断他们的样本是否像随机样本那样具有代表性。

2.1.6 关于相关的严重警告

我们很容易倾向于将相关证据作为因果关系的证据。当然，两件事情可以有关联，却不属于因果关系，造成这种情形的原因可能是多种多样的。

比如说，在对奶酪和奶酪蛋糕的态度之间的这种强相关之下，你也许会得出结论说：一个人对奶酪的喜爱导致他喜欢奶酪蛋糕，因为奶酪蛋糕里有奶酪。但我们也可进行非因果解释：喜爱奶酪的人之所以同时喜欢奶酪蛋糕，也许是因为他们喜欢各种软滑可口的食物。



2.2 相关图表

不论何时，只要发现并且定义了两个变量间的关系，我们就能用其中一个变量去预测另一个变量。画一条回归线，你就可以描绘出相关并做出预测。

假设你刚被任命为区域副经理，负责堪萨斯东北部太阳花湖滨黄金带面积为100 00平方英尺店面的冰淇淋销售。恭喜你！你肩负很多责任，需要做出很多关于如何最大化利润的战略决策。你面临的一个两难问题是：是否需要开门营业。店铺一开门必定要消耗金钱和资源，如果你那天没卖出多少冰淇淋甜筒，可能都不值得打开你那油漆鲜艳的胶合板售货窗口。

要是有某种方法，能神奇地知道任意给定日子的经营状况，那就好了。作为一名统计学爱好者，你认为一定有一种科学的方法，无须通过实际开张、测试当天销售情况，就能估测一天能卖出多少甜筒。你运气不错。的确有一种办法能通过其他信息来估计某个变量（如冰淇淋销量）的分数或值。

关键是你所利用的其他信息必须来自和被关注变量有关联的变量。通过将已知天数里两个变量之间的关系画成一条线，你可以把这条线看作延伸到未来（或是过去），从而预测未知时间点会发生什么。这样的图表工具叫做*回归线*。

2.2.1　勾画未来

善于观察的人总能在变量间发现相关性[Hack #11]。然而，知道存在相关性的好处不仅限于描述性统计。

想象你有太阳花湖畔活动的相关数据。除了其他因素，你已经收集了前任区域副经理任期内的冰淇淋销量（用冰淇淋甜筒数表示）和每天的最高气温（用华氏温度来表示）。代表温度和对冰淇淋购买欲之间相关性的相关系数应该是正向的并且非常大。也就是说，当温度升高时，销量很可能随之增加。

直觉上来说，依据过往经验，你看着温度计，就能大概知道那天冰淇淋摊点的忙碌程度。只要你知道两个变量之间存在正向或负向的关系，可以合理地说，知道一个变量的分数你就能大致了解另外一个变量的分数。

如果你发现两个变量之间存在类似的关系，便可合理假定这两个变量间的关系是线性的。换句话说，如果你画一张图，将一个变量的所有可能值都放在X轴上（沿着底部的水平线），把另一个变量的所有可能值都放在Y轴上（沿着边的垂直线），然后画出每一对分数在象限中的对应点，结果是那些点基本呈直线分布。

2.2.2　连接这些点

图2-1展示了一种用图形来表示湖滨气温和冰淇淋销量间关系的方法。

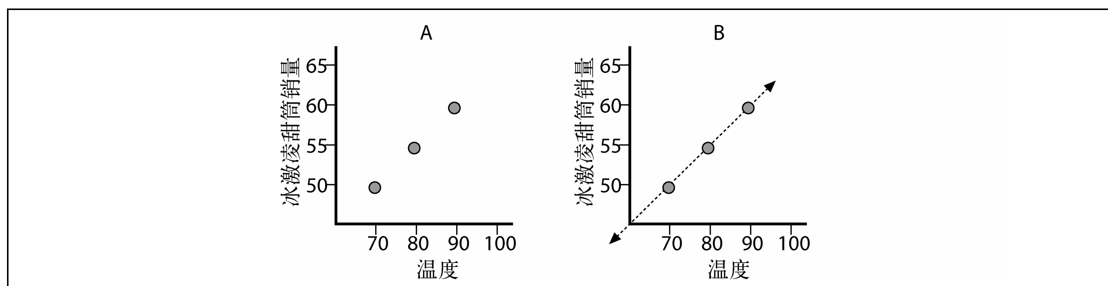


图2-1：销量和温度的线性关系

图A基于你收集到的历史信息，用点来体现两个变量的值。例如，最低点意味着当温度为华氏70度时，卖出了50个冰淇淋甜筒。在90度，卖出了60个冰淇淋甜筒。我们看到一个非常明显的模式，二者间的关系在图上看起来像一条直线。温度每升高10华氏度，甜筒销量就增加5个。温度每升高1华氏度，甜筒销量便相应增加1/2个。图B基于这个规则画了一条线。该线将每个点贯穿起来。

在图2-1中，分析图B能让我们初步认识到回归方程的强大功用。这条线包含了未进行数据取样的区域。例如，我们没有温度是100华氏度的数据。但是，有了回归方程，我们便可估计可能的销量。如果我们在100度标记处找到直线上对应的点，那这个点看起来和65个甜筒的标记相匹配。使用这个回归方程，可以估计在华氏100度的天气里，会卖出65个冰淇淋甜筒。我们同样可以估计较低温度下的情况。图2-1表明，在华氏60度的天气里，会卖出45个甜筒。

2.2.3 玩“如果–怎样”游戏

温度和甜筒销量间的关系可以用数学表达式来表示。以下给出的是图2-1中图A和图B的数据。

温　度	冰淇淋甜筒销量
70	50
80	55
90	60

那么，让我们看看如何用数字建立描述其相关性的方程。毕竟，回归线是统计工具。注意，如果以70度作为起始点，其对应销量为50个甜筒。当我们将70代入公式，应得到50这个结果。同样，代入80应得到55，代入90应得到60。

我用这些数值尝试不同的可能性，试图摸索出输入值与结果值之间的适当数学关系式。我注意到，“冰淇淋销量”的值总是小于温度变量的值，所以我想要一个能够减小温度的方程。线性方程需要一个常量（在每个方程中都使用的某个值）以便产生一条直线，所以我的方程里也需要有个常量。不用反复试验，你也可以把这些数输入到统计程序，比如SPSS，或是电子表格（如excel表）中，生成正确的项。我发现下面这个公式效果不错：

甜筒销量=15+(温度 × 0.50)



从代数上讲，如果你从常量开始，加上一些仅通过基本算术运算（比如乘法）改变的标准量，就能定义出一条表现在图上的直线。

“如果–怎样”是个运用回归线来玩的有趣游戏。在一端输入一个值，就会在另一端得到一个估计值；甚至一些不切实际的情形也能获得答案。在线上放一些疯狂的数值，比如200度，你依然会得到甜筒销量的估计值：115个！

针对这种关系的回归方程，描述的是一条能直观体现该关系的直线。现实中，数据之间的关系很少像我们这个例子中那么清晰。（我们这个虚拟小数据集的相关系数是完美的1.0。）



在统计学里，回归方程使用两组变量分数的相关系数、平均数以及标准差，不考虑数据集中相关的强度。“用一个变量预测另外一个”[Hack #13]说的就是建立回归方程的统计学方法。

2.2.4　生效原理

这类回归估计的准确性有赖于几个重要因素。首先，变量间的相关必须相当大。微小相关产生的点阵式图形根本无法形成直线，通过这些点描绘出的回归线丢失了很多点，并不准确。可惜，在社会科学领域中，我们难以找到太多真正的强相关，所以回归预测往往会产生一定数量的误差。在统计学里，误差是必然存在的。

其次，相关必须至少是线性的。在我们所举的冰淇淋甜筒的例子里，如果在回归线上变量关系发生了质的改变，那么这条回归线就会错失一些数据。幸好，自然界里的相关大多是线性或者接近线性的。

2.2.5　不适用领域

实际的相关不一定是纯粹线性的，但只要基本上属于线性，那么回归分析就相当有效。比如，在我们关于冰淇淋的例子当中，可能温度每上升一度，销量就有所增加。如果在量表的每一处增量都相同，那我们将会看到一种线性关系。但是，在某一温度值上销量骤增也是有可能的。或许，一旦湖畔的气温超过华氏90度，人们就会蜂拥去买甜筒冰淇淋，让自己凉快凉快。

图2-2中的图C和图D显示了相关并非纯线性时，将是怎样的情况。

按照线性回归的要求，回归方程总是生成一条直线，在这种情况下，图中两个点正好落在直线上，但还有一个点不在直线上。通过画相关图来解释数据，这条线完成得很好，但因为相关不

是线性的，所以回归方程产生了一些误差。

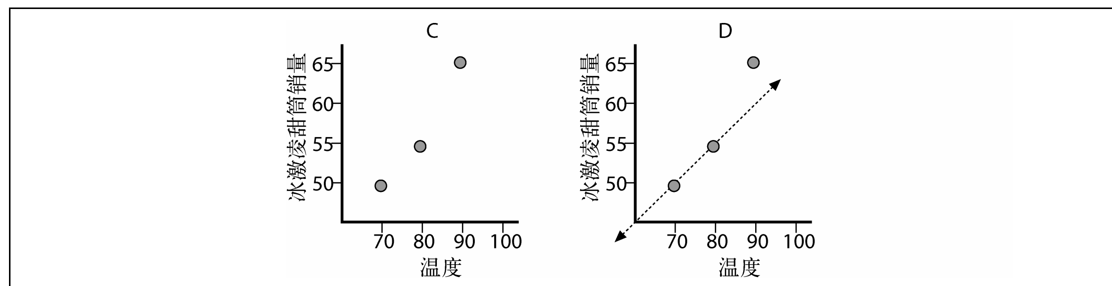


图2-2：非线性相关



2.3　用一个变量预测另一个变量

简单线性回归是一个强大的工具，用来测量你看不见的事物或预测尚未发生事件的结果。有了统计学这个特殊朋友帮忙，你能通过观察某个人在一个变量上的表现，来精确估计他在另一个变量上的得分。

无论在社会科学领域还是其他领域，专家们往往需要预测一个人在某项任务上的表现或是在某个变量上的得分，却无法直接测量这些关键变量。比如，在大学进行录取决策时，这是一个普遍需求。招生委员会想要预测学生入学后的表现（平均绩点或毕业时的总绩点），但是，因为学生并没有实际入学，招生委员会必须利用当前掌握的信息去推测未来。

学校总是用标准化的高考分数作为估测学生未来学业表现的一个指标。假设一所小型大学决定使用美国大学入学考试（ACT）的分数作为学生第一年期末平均绩点（GPA）的一个预测指标。招生委员会回顾了几百个大一学生的GPA以及他们的ACT分数。让他们高兴的是，他们发现这两个变量间存在中等程度相关：相关系数是0.55。

相关系数是两个变量间线性相关的强度度量[Hack #11]，相关系数为0.55表明相关强度相当大。这是一个好消息，因为上述两个变量间相关的存在，使得ACT分数成为估计GPA分数的一个很好的候选指标。

简单线性回归是这样一种方法，它可以生成用来“烹制”预测未来魔法方程的所有数值。这方法生成了一条回归线，画出这条回归线，我们就能判断未来情况如何[Hack #12]。不过，只要有了这个方程式，我们就不再需要通过实际作图去预测。

2.3.1　烹饪方程式

首先，请仔细阅读创造公式的“食谱”（参考“回归方程食谱”），然后我们来看如何用它处理真实数据。你可以把这个食谱剪下来放到厨房抽屉里。

回归方程食谱

配料

相关变量的样本数据2个：

- 效标变量（你想要预测的变量）1个
- 预测变量（用作预测指标的变量）1个

两变量之间的相关系数1个

样本平均值2个

样本标准差2个

容器

空方程式一个，形式如下：

效标=常量+(预测变量×权重)

操作方法

计算用于与预测变量相乘的权重：

权重=相关系数×效标标准差/预测标准差

计算常量：

常量=效标平均数–(权重×预测平均数)

把刚才算出的常量和权重填入空的回归方程。

适合的对象

每个对估计假设结果感兴趣的人

回归方程还需要另外两样配料，即两个变量的平均数和标准差。以下是我们例子中的统计值：

变　　量	平均数	标准差
ACT分数	20.10	2.38
GPA	2.98	0.68



不妨回顾“仅用两个数字描述世界”[Hack #2]，复习平均数和标准差的概念。

招生委员会通过这些信息建立了一个回归方程。结果是，由于所有的申请函都寄到了招生委员会办公室，工作人员能够把学生的ACT分数输入回归方程并预测他的GPA。我们来看一个例子，计算回归方程的各项：

$$权重=相关系数\times效标标准差/预测标准差$$

$$权重=0.55\frac{0.68}{2.38} \qquad 权重=0.55(0.29) \qquad 权重=0.16$$

$$常量=效标平均数-(权重\times预测平均数)$$

$$2.98-(0.16\times20.10)=2.98-3.22=-0.24$$

我们把所有信息代入回归方程，便得到通过ACT分数预测GPA分数的公式：

$$效标=常量+(预测变量\times权重)$$

$$预测GPA=-0.24+(ACT分数\times0.16)$$



注意这个例子中常量是一个负值。那没有关系。

2.3.2　预测分数

还是那个大学录取的例子，想象招生委员会接到两份申请。一位申请者名叫梅利莎，ACT分数是26分。另一位申请者布鲁斯的ACT分数是14分。

使用我们刚建立的回归方程运算可知，这两人最终的平均绩点会有两个不同的结果：

- 梅利莎

- 预测的GPA=−0.24+(26 × 0.16)
- 预测的GPA=−0.24+4.16
- 预测的GPA=3.90

- 布鲁斯

- 预测的GPA=−0.24+(14 × 0.16)
- 预测的GPA=−0.24+2.24
- 预测的GPA=2.00

站在布鲁斯的立场，我希望这所大学的招生名额不止一个。



本例中的两个变量，即ACT和GPA分数，有着不同的度量尺度：ACT分数通常介于1~36，而GPA分数介于0~4.0。相关分析的一个奇妙之处就是，变量的度量尺度可以不同，这没有关系。预测的结果不知怎么就能符合效标变量的度量尺度。听起来有点怪异，是吧？

2.3.3　生效原理

当两个变量彼此相关，它们提供的信息有重叠之处，就如同二者共享信息一样。统计学家有时用相关信息来讨论变量共享变异。

如果一个变量的变异能部分地被另一个变量的变异所解释，那就可以理解，聪明的数学家何以能用一个相关变量去估计另一个变量上平均值的变异（或是距平均值的距离）。他们可能需要用一些数字代表变量的平均值和变异性，用另一个值代表信息重叠度。我们的回归方程包含平均值、标准差和相关系数，这样就使用了上述所有信息。

2.3.4　其他生效领域

回归不仅用在预测上，在回答研究问题上也非常有用。有时候，科学家只想了解一个变量，弄清它的作用原理或在总体中是如何分布的。他们可以通过查看该变量如何与（他们更了解的）另一个变量发生关联来达到目的。



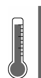
统计学家口中的简单线性回归之所以被称为“**简单**”（simple），不是因为它很容易（easy），而是因为它只用到一个预测变量。这种简单是相对于**复杂**而言的。现实生活中，类似于我们所举例子的预测用到的变量总是不止一个，而是很多。使用多个预测变量来预测效标变量的方法叫做多元回归[Hack #14]。

2.3.5　不适用领域

在三种情况下预测会出现错误。第一，如果两变量间的相关不完美，预测也不会非常准确。鉴于在预测变量和效标变量之间几乎从来不存在真正的大相关，更不用说完美的1.0相关，所以现实世界里回归的应用结果错误百出。尽管如此，只有存在任何相关，预测就比盲目猜测更准确。你可以通过标准误差估计[Hack #18]来算出误差的大小。

第二，线性回归假定关系是线性的。这在“相关图表”[Hack #12]里详细讨论过，但如果相关的强度在分数分布区间的不同点上存在变化，那么在一些情况下回归预测会产生很大的误差。

第三，如果最初收集的用于确定回归方程中各个值的数据不具备对未来数据的代表性，那么预测结果就会出错。比如，在大学录取的例子中，如果一个申请者的ACT分数是36分，那么预测的GPA值将为5.52分。这是一个不可能的值，甚至都不匹配GPA的度量尺度，GPA最大值为4.0。因为用来建立预测方程所用的过往数据极少或根本没包含ACT为36的值，以致回归方程无法处理如此之高的分数。



HACK #14

2.4　用多个变量预测单个变量

任何统计黑客都可掌握预测未来和看到不可见事物的超能力，只要他们觉得这种能力有价值。统计学家总是用一个变量预测另外一个，以此来回答问题，并用相关信息来解决问题。但为了更准确地进行预测，可以使用多元回归的方法，将不同的预测变量结合在同一个回归方程里。

“相关图表”[Hack #12]中讨论了回归线在预测方面的有用性。利用这些方法，行政人员和统计研究人员能够预测尚未发生的评估表现，理解变量，建立关于这些变量间相关性的理论。他们只用一个预测变量就能完成这些技巧。

“用一个变量预测另外一个变量”[Hack #13]中展示了大学招生录取时遇到的一个难题：他们想要录取未来学业出色的学生，所以他们尝试预测学生的未来表现。这条Hack所采用的方法是用一个变量（标准测试分数）去估计未来变量的表现（大学成绩）。

现实世界中，研究人员经常想要利用多个变量中所发现的信息（而不仅仅是一个变量）来预测或估计分数。如果追求更高的准确度，科学家们会尝试寻找多个看起来都和效标变量（你想要预测的变量）相关的变量。他们利用所有这些信息生成一个*多元回归方程*。

2.4.1　选择预测变量

在深入探讨本条Hack前，你或许应当阅读或者重温“用一个变量预测另外一个变量”[Hack #13]，只是为了回顾一下手头的问题以及回归法是如何解决该问题的。下面是我们在[Hack #13]中建立的、使用ACT分数作为单一预测变量的方程式，用以估计未来大学录取情况：

$$\text{预测的GPA} = -0.24 + (\text{ACT分数} \times 0.16)$$

这个单一预测变量生成了一个回归方程，结果ACT与GPA相关度为0.55。非常好，非常准确，但它还可以更好。

想象一下，假设该大学的管理者觉得自己刚建立的回归线或回归方程所得结果还不够准确，想要做得更好。如果他能找到更多的和大学成绩相关的变量，就能获得一个更准确的结果。不妨假设我们的业余统计学家发现了另外两个和大学表现相关的预测变量：

- 态度度量
- 论文质量

态度调研分数可能是由大学收集的（分值范围为20~100分），人们发现它与学生未来的GPA具有某种相关。此外，个人论文得分（分值范围为1~5分）也可能和大学GPA相关，或许能够包含在多元回归方程里。

2.4.2　建立多元回归方程

我们先大体看看回归方程的抽象形式，然后再将这一工具应用于手头的任务。以下是仅用一个预测变量的基本回归方程：

效标变量=常量+(预测变量 × 权重)

如果你想利用更多的信息，可以扩展这个方程，使其包含更多的预测变量。下面的回归方程包含三个预测变量，但你还能扩展该方程，将更多的变量纳入其中。

效标变量 = 常量+

(预测变量1 × 权重1) +

(预测变量2 × 权重2) +

(预测变量3 × 权重3)

每个预测变量都有其对应的权重，其大小是通过基于预测变量和效标变量相关的统计学公式确定的。具体计算过程有些复杂，在此就不予展示了——小意思，不用谢我。在现实中构建回归方程时，人们几乎总是用计算机来生成多元回归方程。



本书中的许多运算都是运用统计软件SPSS完成的：我把数据（通常是虚拟的）输入SPSS数据文件中，从而得出结果。微软的Excel也一个进行简单统计分析的便利工具。

使用我们能找到的包含三个与效标变量相关的预测变量（各预测变量之间也存在某些相关）的实际数据，我们可以生成一个回归方程式，其值如下：

预测GPA = 3.01+

(ACT分数 × 0.02) +

(态度分 × 0.007)+

(论文分 × 0.025)

我在我的电脑上使用这些想象的数据计算出上述权重。总体上，该方程可以很好地预测大学GPA，在观测的GPA分值和预测的GPA分值间找到了0.80的相关，大大高于我们使用单一预测变量产生的0.55相关。



当我们在模型（对一组变量以及变量如何相关的描述）中加入另外两个预测变量，即态度测量和论文分数，ACT的权重即随之发生改变。这是因为对每个变量都用部分相关替代了一对一的相关。此外，常量也变了。2.4.2节会对此加以讨论。

2.4.3 作出预测并理解相关

为了估计某个学生未来在大学期间的学业表现，校方管理人员把该生在每个预测变量上的分数输入回归方程，然后将每个变量分数乘以其对应的权重再加上常量，所得的值就是对该生未来学业表现的最佳估计。当然，这可能不完全正确（实际上这种可能性很大），但总比没有任何信息要好。



如果你不掌握任何信息，只能凭空估计一个学生在大学里的学业表现，你应该估计他的分数相当于平均GPA分数，不管你们学校的平均分是多少。

假如你不仅想要预测未来，还想透彻理解预测变量和效标变量之间的关系，那又如何？你的目的可能是想建立一个更有效的公式，其中无须包含众多无用的变量；也可能是想建立一种用于理解这个世界的理论——你这个疯狂的科学家！问题在于，你很难做到仅看权重就知道每个预测变量的独立贡献。

在多元回归方程里，每个变量的权重是和每个变量实际的分数区间范围成比例的。这样就很难通过对比各个预测变量来判断哪个在预测效标变量时提供的信息最多。对比这些原始的权重可能会产生误导，因为一个变量的权重较小，可能只是因为它的度量尺度较大。

例如，我们来对比ACT分数的权重和态度分数的权重：ACT分数的权重是0.02，比态度的权重0.007要大，但不要误以为ACT分数在预测GPA时比态度分数更重要。记住，GPA分数范围是1.0到4.0左右，而态度分数范围是20分到100分。与较大的ACT分数权重相较而言，较小的态度权重却造成效标变量产生更大的变动。

多元回归分析的计算机程序结果总是显示为表2-4那样格式的信息。

表2-4：多元回归结果

效　标	非标准权重	标准权重
常量	3.01	----
ACT分数	0.02	0.321
态度分数	0.007	0.603
论文分数	0.025	0.156

在确认关键预测变量和对比每个预测变量在估计效标变量的独特贡献时，表2-4的第三列比“非标准权重”列的值更有用。



标准权重就是原始数据转换成*Z*分数[Hack #26]后得到的权重，*Z*分数是用标准差来表示每个原始数据和平均数的距离。

标准权重将所有预测变量纳入了同样的度量尺度。这样一来，可以公平地对各个预测变量对效标变量的相对重叠部分加以对比和理解。比如，运用这些数据，或许可以适当地说，态度对大学GPA的解释量是ACT表现解释量的两倍，因为态度的标准权重是0.603，大约是ACT分数权重（0.321）的两倍。

2.4.4　生效原理

多元线性回归在预测结果时表现得比简单线性回归要好，是因为多元回归使用了一点额外的信息，来计算每个预测变量的实际权重。多元回归知道每个变量和其他变量之间的相关，并用这种相关去生成更准确的权重。

这点复杂性是有必要的，因为如果预测变量之间存在相关，它们就会共享一些信息。如果它们互相相关，那它们就不是真正的独立预测源。为了使得回归方程尽可能准确，统计学方法移除了方程中每个变量互相共享的信息。这样就产生了从不同角度对效标的独立预测变量，生成了尽可能准确的预测。



想象两个预测变量，二者间呈完美相关，相关系数是1.00。在同一个回归方程里使用这两个变量，并不会比只用一个（不管是哪个）要准确。引申可知，两个预测变量间的任何重叠（比如，两个预测变量间任何大于或小于0.00的相关）都是冗余信息。

图2-3说明了使用多个独立信息来源去估计一个效标分数的情形。

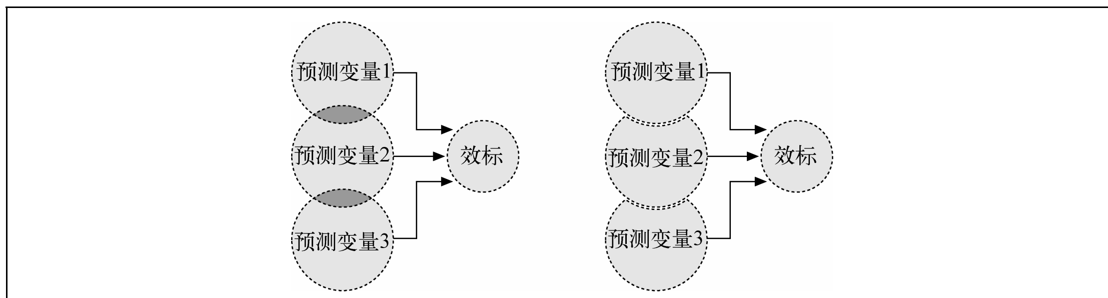


图2-3：多元回归中的多个预测变量

在多元回归中，用来决定每个预测变量权重的相关信息，不是预测变量和效标变量的一对一相关，而是当所有预测变量之间的重叠被移除后，预测变量和效标变量之间的相关。

这个过程产生了有点不同于实际测量变量的预测变量。通过统计学上的移除（或是控制）变量间的共享信息，预测变量在概念上变得不同于以往。正如图2-3所示，现在它们是有着不同“形状”的独立变量。这些改变后的变量和效标变量的相关被用来生成权重。

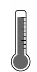


当所有冗余信息从预测变量上被统计移除后，预测变量和效标变量之间的相关被称作**部分相关**。部分相关是在预测变量和其他变量不相关的情况下，每个预测变量和效标变量之间一对一的相关。

2.4.5　其他生效领域

在现实世界中，多元回归每天实实在在地被人们使用着，其原因有二。首先，多元回归使得预测方程的构建成为可能，所以人们能够用已有的群组分数去估计另一个不在眼前的变量的分数（因为它尚未发生或是由于某些原因无法轻易测量）。这就是多元回归这种统计工具在应用科学领域解决问题的原理。

多元回归同样能检验一组变量对其他变量的独立贡献。它让我们看到哪里有变量间的信息重叠，并建立理论去理解或解释这种重叠。这是多元回归这种统计工具在基础科学领域解决问题的原理。



HACK #15

2.5　识别非预期结果

你怎么知道你的观测值是否正确，或者是否存在偏差？你怎么知道什么时候某件事发生的概率可能比原本应该发生的概率大或小？你可以凭借灵活的单因素方差检验来确切地获知以上问题的答案。

在科学领域，最古老的观测研究包括清点人员、动物和物件的数目。

- 这条船上有多少人？
- 翅膀上带绿点的蝴蝶比例是多少？

随着推断统计走向成熟，问题变得更加具体。

- 1812年伦敦出生的男孩和女孩数量相同吗？
- 一天中不同时段发生的罪案数量相同吗？

这些情境中的研究问题是“他们/它们的数量相同吗”。（或者至少是，他们/它们的数量是否足够接近以至于任何波动皆可能出于偶然？）不等同分布的意思就是有事情在发生。它无法回答实际在发生什么的问题。但这是一个开端，一个很好的初始问题。

你是否注意到某种异常情况似乎正在发生，但不确定那是否只是你的想象？在本地社区商场购物的嬉皮士是否异常增多，超出了偶然几率？如果答案是肯定的，而且你很想遇见嬉皮士，那你就该在商场附近多转转。

对于商家和服务业人士来说，确认哪里有最大的需求非常重要。观测数据能够用来解决这个

问题。甚至只在日常生活中，我们都有自己的基于观察的信念（有可能有偏差）。我已经注意到社区商店里有很多嬉皮士，但这或许是因为我当时特别留意了嬉皮士。那个地方的嬉皮士确实比平常多吗？比普通人更多？

这类问题可以借助一种统计方法来回答，这种方法适用于查看多个类别中的每一类所包含的“事物”数量较之正常状态下随机发现的数量是否有所不同。这方法叫做单因素卡方检验。



这种统计分析之所以称作“卡方检验”，是因为它用符号**C**（即希腊字母**X**，读作/kai/）来表示生成的临界值。计算中所需的值都是平方值，所以我们将其统称为**卡方**（英文为**chi-square**或**chi-squared**）。

2.5.1　判断是否有异常情况

假设你负责制定你所在城镇的警察巡逻时间表。问题在于，你不知道是否该对每一班安排相同数量的警员，或许特定班次需要应对的罪案数量更多？如果某个班次可能会忙些，那你应该在这个时段分配更多的警员。当然，在该时段分配更多警员，加强巡视，也能起到抑制犯罪的效果。

下面的例子以虚构的数据表现了三个时段内犯罪事件发生的情况。假设这些数据取自30天的时间，你想要利用它们为来年做计划。表内数字代表三个时段中每一时段发生的犯罪数量。

午夜12点至早晨8点	早晨8点至下午4点	下午4点至午夜12点	总　计
120	90	90	300

无疑，看起来更多的犯罪发生在深夜。仅通过观察，我们就可能得出结论说，深夜里的罪案更多。但是，那也许仅在我们的样本中如此，而在总体数据中并不存在这样的差异。

2.5.2　计算卡方

我们能计算这个数据的卡方值。如果卡方值的确很大，那就说明深夜时段的罪案数量（120次）异乎寻常地大于另外两个时段的罪案数量。多大才算“的确很大”，这是个很重要的问题，我们将在本节稍后部分加以讨论。



可以按以下脉络来理解我们即将做的分析：如果24小时内总共发生了300次犯罪，那么我们可以预期一天内三个相等时间段内各有33.3%的犯罪，也就是说每个时段有100次犯罪发生。如果任意一个时段内的罪案数量超过或少于100次，就意味着有异常情况。也许时间段对犯罪的发生有影响。当然，也可能有些偶然的波动，但期望频次和实际频次的差异越大，这种差异仅仅是出于偶然的可能性就越小。

以下是卡方计算公式：

$$卡方=\sum\frac{\left(观测频次-期望频次\right)^2}{期望频次}$$

Σ表示求和或是把它后面的各项相加。

让我们计算这个数据的卡方。每个类别的观测频次是给定的。每个单元的期望频次是300除以类别数量3，即100：

$$\frac{(120-100)^2}{100}+\frac{(90-100)^2}{100}+\frac{(90-100)^2}{100}$$

$$\frac{(20)^2}{100}+\frac{(-10)^2}{100}+\frac{(-10)^2}{100}=\frac{400}{100}+\frac{100}{100}+\frac{100}{100}=4+1+1=6$$

这个数据的卡方值是6。很好。现在该做什么？6是大是小还是其他？卡方值大小为6是偶然的吗？

2.5.3 判断卡方值是否“的确大”

就像相关系数[Hack #11]、t检验[Hack #17]、比例等其他一切统计方法一样，统计学家已经标定了卡方的分布。换句话说，我们知道不同大小的卡方偶然出现的可能性。发现特别巨大的卡方值的可能性，取决于类别的数量。

表2-5为一张理论上超级庞大的表格的一部分，表示为了有95%把握（显著水平=0.05）必须达到的卡方值。如果卡方值没那么大，说明只是样本中的偶然波动导致的。我们知道这些临界值有5%或更小的几率出现，因为卡方值就像有序的统计世界中几乎所有其他事物一样，有着一个已知的分布，即一个特定值出现可能性的已知集合。像正态曲线一样，卡方分布也具有明确的定义[Hack #23]。

表2-5：显著性为0.05的卡方临界值

两个类别	三个类别	四个类别	五个类别
3.84	5.99	7.82	9.49

我们的卡方值是6，大于三个类别对应的临界值（5.99）。这意味着存在某些异常的情况，所以我会强调它。虽然这里是以犯罪发生率问题作为一个具体事例，但我使用的描述方式通用于所有在0.05显著水平的统计发现。



如果在总体中，一天中三个时段的犯罪数量并无差异，你也可能随机抽取到有差异的样本，产生的卡方值为6或更大，但这种情形发生的概率低于5%。

那么，看来我们可以合理地得出结论，总体中一天内不同时段的犯罪频次存在差异。因为这些差异是“真实的”，所以有理由在此基础上安排全年的警察巡逻计划。

2.5.4 生效原理

卡方分析的数据表示方式能使每类事物的观测数量和每类事物的期望数量相比较。“每类事物的期望数量”总是被定义为一个相等的数。如果没有异常情况（比如，类别之间没有差异），那么可以预期各个类别中事物的数量相等。

卡方适用于类别数据。实质上，每个类别的期望值和观测值之间的差异都会被计算。将差异和期望频次加以对比（作为一种标准化所有差异的方法），然后把所有的比例数字相加。相加得到的结果是它偶然出现的可能性。数字越大，用偶然性单独解释事情的可能性越低。存在一个已知的分布（每个可能卡方值对应的概率），通过表格（或计算机）将特定的概率派给每个卡方值。

如果类别数大于或等于2，研究人员又想知道这些类别中的实际分布情况是否与偶然出现的相同，那么卡方就是一个合适的检验方法。实际检出值是研究者预期发现和实际发生之间的差异。

卡方检验的使用框架是：研究者怀有某种预期，以此方法来检验观测数据是否与该预期相符。这是一个简单的模型检验形式。研究者有一个信念系统，以某些模型或假说（认为现实世界的运转方式应当如何）的形式存在。随后他就观察世界（收集数据）并将他的观测结果和模型加以对比。如果数据符合模型，便对假设形成支持。因此，卡方检验被视为一种*拟合优度统计检验法*。它回答的是数据在多大程度上与模型相符的问题。



有些统计教科书把**单因素卡方检验**称作**单样本卡方检验**，所以不要把它们搞混了。但是，难道你还有必要去读别的统计书吗？

统计学家了解观测频次较之期望频次可能出现的常态波动的大小。掌握了这个信息，他们就能计算观测值与预期值之间出现任何偏离的可能性究竟是出于偶然还是缘自其他因素的影响。

2.5.5 其他生效领域

卡方检验作为一种统计学方法虽然简单而古老（大约80年历史——在统计学领域已经算是“古老”了），但它对于很多统计问题的解决都非常有用，从测量标准较低的到非常高级的统计方法都是如此，惊讶吧！因为它是一种相当直接的模型检验（或“拟合优度检验”）方法，所以卡方检验被用作复杂的相关分析以及测量诊断的一部分。

卡方分析用来检验复杂的理论模型是否真的与现实世界的数据相符，这些理论模型是变量间相关性的详细说明。如果现实数据距这些模型的预期偏离太多，那么我们可以下结论说模型为弱。卡方显著性是显示“过多”偏离的判据。

比如，测试研发人员若关注测项偏差——某一测项对于一个可识别群组（如种族、性别等）的作用可能不同于对另一群组的作用，他们会检查答案选项的模式是否符合某种预期，而不考虑是哪组产生的数据。卡方检验分析是对预期和实际测试表现加以对比。

2.5.6 参阅

“识别非预期相关”[Hack #16]。



2.6 识别非预期相关

如果你想弄清自己观测到的两变量之间的相关是否为真，有很多统计方法可以选用。但是当你对这些变量使用类别测量法进行准确性不太高的测量时，就会出现一个问题。其解决办法是采用两因素卡方检验，这种方法除了其他一些用处以外，还可用来对初识者的特征做暂无事实根据的假设。

在“识别非预期结果”[Hack #15]中，我们采用了单因素卡方检验，依据一天中不同时段犯罪数是否相同而制定警察巡逻班次。这种方法对于解决如下情况的分析问题大有效力。

- 数据属于分类测量的范畴（如性别、党派、种族等）。
- 你想要判断某些特定类别中分数的频次是否高于其随机出现的频次。

当你对两类变量是否互相关联感到好奇时，你会遇到另外一个常见的分析性问题。类别变量间的相关能够用方便的两因素卡方检验来考察。



如果两个变量属于区间测量的范畴（在一个连续体上可能存在多个分数），相关系数[Hack #11]是最好的工具，但这种工具在类别测量方面表现并不出色。

我们一直在对以上类型的变量间相关做出假设。我们给人划分类型的很多常见刻板印象其实就是无形中对这些关系做出假设。下面这些你可能抱有的假设，就隐含着类别变量之间的相关性。

- 教授们总是心不在焉的。
- 程序员玩《龙与地下城》（*Dungeons and Dragons*）游戏。
- 本书作者是喜欢收集漫画的成年人。
- 教授们总是心不在焉的。

如果你在聚会上碰到一名程序员，并且对程序员群体持有上述刻板印象，你可能会假定他熟悉20面骰子游戏。但是，如果你错了，就会使双方的交谈陷入尴尬。所以最好还是先了解你所着眼的类别变量之间是否真的存在上述相关。计算两因素卡方能解决这个问题，并且能够证实或质疑这些关于人的假设。

单因素卡方回顾

卡方检验是在如下框架下使用的：研究者事先抱有某种预期，想看看观测数据是否与这种预期相符。统计学家了解观测频次较之期望频次可能出现的常态波动的大小。掌握了这个信息，他们就能判断观测值与预期值之间出现任何偏离的可能性究竟是出于偶然，还是缘自其他因素的影响。这些分析的原始数据总是某个变量类别中的人数（或者频次）。

以下是计算卡方的通用公式：

$$卡方 = \sum \frac{(观测频次 - 期望频次)^2}{期望频次}$$

Σ表示对它后面的数求和。卡方值越大，结果随机出现的可能性就越小。

2.6.1　回答相关性问题

单因素卡方分析的是单一类别变量，而两因素卡方分析的是两个类别变量之间的相关。二者的内在原理是一样的：将每一类或组合类的期望频次与实际频次加以对比。如果差异之和达到了一个很大的数，那么就有其他因素作用的影响。

这儿有一个我们可能很想找到答案的类别相关问题。它和其他有待探究的刻板印象问题大同小异。

女性更倾向于加入民主党还是共和党？

你心里可能对此已经抱有某种假定，但你要如何去检验这样一个假定的准确性呢？

1. 执行预备分析

首先来看表2-6的例子，其中显示了一组单一分类变量的频次数据。这些数据是虚构的，但和公开发表的研究结果一致，通常发现共和党人士大多是男性，而女性倾向于认同民主党。

表2-6：共和党假设样本

男性	女性
45	30

在这个75个随机抽取的共和党人样本中，45名是男性，30名是女性。即60%为男性，40%为

女性。我们能否这么下结论，认为共和党的成员通常是男性多于女性？否则的话，我们就会预期样本中男女各占50%。



单因素卡方检验能知道共和党员中男性是否多于女性，但那不是本条Hack要探讨的问题。

然而这不是我们的研究问题。

2. 计算两因素卡方

我们开头的问题只包括了共和党，所以在第一次分析中，党派看起来像是一个变量，但它其实只是对总体的一个描述；它没发生任何变化。但是我们可以添加另一个类别——比如说民主党——再招募75个被试，这样我们马上就有了两个变量的数据。假设这些数据如表2-7所示。

表2-7：选民的假设样本

党派	男性	女性	总计
共和党	45	30	75
民主党	34	41	75
总计	79	71	150

这里我们有两个分类变量：所属党派和性别。我们可以继续使用单因素分析，分别分析这两行数据。但是，一个更有代表性的问题或许是："党派和性别之间存在相关吗？"



问："党派和性别之间存在相关吗？"
答：这让我想起了大一的时候。
（哈！这种笑话我这儿多的是！本周内我都会在这儿。大家晚安！[1]）

为了计算期望频次和观测频次之间差异性的标准测量，我们使用和单因素卡方分析一样的公式。正如"识别非预期结果"[Hack #15]中所示，我们首先要加总每个单元格（表上的每一格）内的预期和观测频次之差。

我们对两因素卡方做同样的运算。每个单元格的期望频次等于单元格所在行的人数乘以单元格所在列的人数，然后除以样本总数。使用表2-7的数据，对期望频次的计算展示在表2-8里。

注1：20世纪美国著名谐星杰米·杜兰特（Jimmy Durante）的一句口头禅。——译者注

表2-8：两因素卡方分析期望频次

党派	男性	女性
共和党	（75×79）/150=39.5	$(75\times71)/150=35.5$
民主党	（75×79）/150=39.5	$(75\times71)/150=35.5$

所以，两因素卡方的计算如下所示：

$$卡方=\frac{(45-39.5)^2}{39.5}+\frac{(34-39.5)^2}{39.5}+\frac{(30-35.5)^2}{35.5}+\frac{(41-35.5)^2}{35.5}$$

$$卡方=\frac{(5.5)^2}{39.5}+\frac{(-5.5)^2}{39.5}+\frac{(-5.5)^2}{35.5}+\frac{(5.5)^2}{35.5}$$

$$卡方=\frac{30.25}{39.5}+\frac{(30.25)}{39.5}+\frac{(30.25)}{35.5}+\frac{(30.25)}{35.5}$$

$$卡方=0.77+0.77+0.85+0.85=3.24$$

3. 判断卡方值是否足够大

统计学家知道2×2表格（就像我们刚才计算的卡方一样）的卡方临界值是3.84。在随机情况下，卡方值大于3.84的几率大约为5%或更少[Hack #15]。

因为我们的卡方值是3.24，小于临界值3.84，于是我们知道这样一个波动随机发生的概率高于5%。这里我们还不能宣称达到统计显著性，因此我们必须下结论说，虽然我们的样本似乎显示所属党派和性别这两个类别变量之间存在某种相关，但这可能是因为取样误差所致。在我们抽样的总体中，可能不存在任何相关。

2.6.2　生效原理

两因素卡方通过观察差异性来回答此类相关性问题。这可能看起来有违直觉，因为大多数统计是通过寻找不同来展示差异，而不是展示相似性。但其中的思维逻辑是：

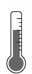
- 如果党派和性别之间不存在相关，那么每个性别群组中共和党人和民主党人应当各占一半；
- 同样，如果党派和性别之间不存在相关，那么每个党派内部的男女成员也应当各占一半；
- 这种双向的等同分布应为随机形成。相对于上述预期的较大偏离表明有外界因素作用的影响。

本项Hack可用于检验我们持有的刻板印象是否正确。当然，在超乎现实世界的科学领域，研究人员还使用这个方法去探索形形色色的复杂问题。

两因素卡方分析有时候称作列联表分析，当你手上有两个类别变量并且想知道一个变量对另一个变量是否有某种依赖时候，这种方法非常有用。本例中的变量只有两个类别，但我们可以以此类推，对多个类别的变量进行分析。它的技术要求会有点复杂，但步骤是一样的。

2.6.3 参阅

“识别非预期结果”[Hack #15]



HACK #17

2.7 比较两组

哪个更好？哪个更多？人与人之间真的有差异吗？诸如此类的定量问题是我们礼节性谈话内容当中的重头戏。如果你想要拿出真实的证据来支持自己关于哪个最好、哪个最多、哪个最少的观点，可以使用一种叫做“t检验”的统计工具来达到目的。

我叔叔弗兰克的脑子里总是充满这样那样的观点。比如，他认为绿色的M&M巧克力豆比蓝色的味道好，他认为女性从来不会收到超速罚单，他认为《脱线家族》（*Brady Bunch*）中的孩子们唱得比《鹧鸪家庭》（*Partridge Family*）中的更好听。还有，他认为格子花呢又回归潮流了。他一天到晚接二连三地抛出那些不成熟的观点。虽然在上述四个问题上我都持不同意见（尤其是说到格子花呢回归潮流这一点，因为它从来都没有退出时尚），但我只能直接说明我的观点来反驳他，除此之外拿不出别的证据。

要是有某种科学的方式来证明我叔叔弗兰克是对还是错，那该有多好！——你无疑能够看出我这句话是在玩弄修辞手法。事实上，可用来检验此类假设的统计方法多得不可胜数。其中一个最简单的工具，其设计目的就是为了检验最简单的声明。如果你想判断两个组别之间是否存在差别，那么独立t检验就是最好的解决办法。

2.7.1 证明弗兰克叔叔是错的（或对的）

为了应用t检验来实际考察弗兰克叔叔的一个理论，我们必须计算出一个t值。假设我打算真正挑战一下弗兰克叔叔，并且收集了一些数据来检验他的观点是否站得住脚。

弗兰克叔叔认为男性收到超速罚单的频次要高于女性。为了检验这个假定，不妨想象我从他的邻居中，随机选取[Hack #19]了两组开车者，每组15人。其中一组是女性，另外一组是男性。假设我问了他们一些问题。结果发现在过去的5年间，男性组平均收到1.71次超速罚单，方差大小为0.71；女性组平均收到1.35次超速罚单，方差大小为0.25。



方差是给定一组数中，总的变异量大小。它是通过找出群组中每个分数和平均分数的距离而计算出来的。将这些距离进行平方并求算术平均数就能得到方差值。

下面是t值的计算方程式：

$$t=\frac{\text{第一组平均数}-\text{第二组平均数}}{\sqrt{\frac{\text{第一组方差}}{\text{第一组样本大小}}+\frac{\text{第二组方差}}{\text{第二组样本大小}}}}$$

t值越大，在你样本群体中发现的任何差异性为随机出现的可能性越低。通常情况下，当t值大于2就足以下结论说，差异不仅存在于你的样本中，也存在于整个总体中。



这里给出的t值计算公式，在两组人数相同时效果最好。当两组样本量不等时，会采用一个类似的对变异信息求平均的公式。

对弗兰克叔叔的观点是否得到支持？为了确定这一点，我们的计算需要用到表2-9的数据。

表2-9：超速罚单t检验数据

	第1组（男性）	第2组（女性）
平均数	1.71	1.35
方差	0.71	0.25
样本量	15	15

如果我们把这些关键值代入前面的公式，就会得到：

$$t=\frac{1.71-1.35}{\sqrt{\frac{0.71}{15}+\frac{0.25}{15}}}$$

于是得出计算结果：

$$t=\frac{0.36}{\sqrt{0.047+0.017}}=\frac{0.36}{\sqrt{0.064}}=\frac{0.36}{0.253}=1.42$$

在这种情况下，我们由0.36的平均差计算出t值大小为1.42。

2.7.2　解释t值

这个1.42大小的t值会是随机发生的吗？换句话说，如果总体中的实际差异为零，从这个总体中抽取的两个样本平均值会有那么大的差异吗？

之前我提到过，若要得出这个结论，通常需要t值为2或者更大。在这种标准下，我们会下结论说，没有证据显示男性的确比女性接到更多的超速罚单。当然，在我们的样本中是这样，如果

我们测量所有人（全部总体），结果可能就不是这样。没有证据显示弗兰克叔叔是对的。虽然这并不等于说他是错的，但依然意味着他的这个论点站不住脚。

但是，统计学是讲究准确性的学科，所以让我们来进一步探究1.42这个值。*t*值具体要达到多大，我们才能下结论说弗兰克叔叔真的是正确的？

依照惯例，如果在某一*t*值水平上随机概率为5%（或更小），该*t*值即被视为足够大。幸运的是，从总体中随机抽得不同*t*值的几率，已经被辛勤的数学家们利用中心极限定理[Hack #2]计算出来了。统计显著性需要的实际*t*值大小，取决于两组样本的总和。表2-10提供了达到0.05统计显著水平需要满足或达到的*t*值。

表2-10：随机出现几率小于5%的*t*值

两组联合样本量	临界*t*值
4	4.30
20	2.10
30	2.05
60	2.00
100	1.99
∞（无穷大）	1.96



对于未包含在表2-10中的样本量，你可以通过估计表中两个*t*值之间的值，来得出你需要满足或达到的粗略的*t*值。同样，该表假定你想要在两个方向中的任一方向确定组间的差异性。它假定你想要知道其中任意一组的平均数是否大于另外一组的平均数。这就是统计学家所称的**双尾检验**，这通常是一种有趣的对比。

查阅表2-10，我们看到*t*值为1.42时，小于30个被试总数的临界值2.05。如果需要有把握地说我们观测到的样本差异不只是出于偶然，那我们需要看到一个大于2.05的*t*值。

2.7.3 生效原理

社会科学家一直在使用这种对比方法。实验设计和准实验设计总是设置两组人群，两组间被认为在这样或那样的方面存在差异。你可能着眼于共和党和民主党之间的差异，或是男孩和女孩之间的差异，或是想看看服用新药的群组中患感冒的人数是否比不服用任何药物的群组更少。

这样的设计会产生两组分数，它们的值总是存在差异，至少在使用的样本间存在差异。研究者（当需要证明弗兰克叔叔是错误的时候，我也算是一个研究者）更感兴趣的是，两组样本所代表的总体之间是否存在差异。



推断性统计的逻辑是：样本分数代表一个更大的总体的分数。如果样本在某个变量上存在差异，那这种差异也许能被反映在它们来自的总体中。还有另一种可能，这种差异也许是缘自取样误差。

t检验回答了这样一个问题：两样本间发现的任何差异究竟是真实的（即，它们很可能存在于样本来自的总体），还是缘自取样误差（即它们很可能只存在于样本中，总体中不存在）。如果样本间的差异太大以至于无法用偶然出现来解释，那么研究人员就能下结论说总体间存在真实的差异。

t检验公式使用了样本分数分布形状的信息。我们需要每组研究变量的平均分数，每组的方差，以及每组的样本量这几个信息。样本平均数提供了对总体平均数的很好估计，方差指示样本平均数可能偏离总体平均数的程度，样本量提供了估计的准确性。两个平均数之间的差异被标准化且用一个t值来表示。

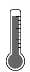


当统计学家谈论真实差异时，他们会说“这两个样本可能来自不同的总体”。而你我以及具体研究者谈论真实差异的方式则可能是“共和党和民主党存在差异”，或“此药物降低了患感冒的几率”。

2.7.4　其他生效领域

数字并不知道自身来自何处。你可以用t检验去检查任意两组数字的差异性，不管它们描述的是人还是物。实际上，t检验的发明，最早为了在啤酒生产中判断整仓谷物的质量。

一名啤酒统计学家（梦想中的职业啊）想要发明一种方法，只需从谷物总体中随机抽取一小部分样本进行查验，而不是检验所有的谷物。剩下的故事就是历史了。所以我们今天可以说，统计研究人员所做的大部分工作的的确确是由啤酒驱动的。



HACK #18

2.8　看清实际错误程度

任何时候你使用统计量来概括观测数据，你都有可能犯错。如果你需要知道自己已经多么接近真相，可以使用标准误差这个工具。

在专业人士当中，或许唯有统计学家不仅自豪地承认自己的答案可能出错，而且会想尽办法精确地告诉你，他们实际上错到什么程度。当你执行一项调查，记录观测数据，或是执行某种类型的实验，你的结果所描述的仅仅是你的样本——你面前的顾客、患者、学生、金鱼或是成片的氪气石。推断性统计利用样本算出的值来估计样本代表的总体中相应值的大小。比如，根据样本中的平均数可以很好地估计出总体的平均数。问题是你要知道是否应当信任你的结果。

2.8.1 校准误差并计算精确性

一个样本的平均数不太可能和总体平均数完全一样，但很可能接近总体平均数。如果你想要知道自己的错误程度，那么能用标准误差来校正你的准确性。通过平均数的标准误差，可以大致估计出根据样本得出的估计平均数和实际总体平均数之间的差距。



1.6节讨论了如何在测量中使用标准误差。计算标准误差让你知道你的测试分数和典型表现水平有多接近。正如测量使得我们能够计算个体观测分数附近95%的置信区间，统计学家通常针对众多样本值计算其附近95%的置信区间。

2

幸好，对于任何想了解统计发现和潜在真相之间差距有多远的人来说，每个流行的统计方法都会提供一个标准误差。在介绍完下面的基本概念后，本节接下来会解释如何运用这些标准误差。

- ❑ 描述性统计中的平均数标准误差。
- ❑ 调查取样中的比例标准误差。
- ❑ 回归中的估计标准误差。



在取样时，中心极限定理[Hack #2]是了解我们错误程度的关键工具，因为它提供了计算标准误差的公式并且提示所有样本概括值均呈正态分布。

利用标准误差来核实统计分析结果的准确性，常用的方式有三种。选择哪个特定工具，取决于你是否想知道自己在多大程度上接近正确的估计。

- ❑ 某个变量的总体平均分（例如，无任期保障的大学教授的平均工资）。
- ❑ 总体中拥有某个特征的成员所占比例（例如，哪些人会投票支持我叔叔弗兰克担任捕狗员）。
- ❑ 未来的表现（比如，你那只受过选择题答题训练的宠物猴可能获得的大学GPA成绩）。

2.8.2 平均数估计

样本平均数作为总体平均数估计值的准确性是以样本量为基础的。其计算公式如下：

$$\text{平均数标准误差} = \frac{\text{标准差}}{\sqrt{\text{样本大小}}}$$

随着样本量的增加，样本平均数越来接近于真实的总体平均数。如果你将样本量想象成独立观测数量的话，这个现象就能讲得通了：你对一件事物的观测次数越多，你的描述就越准确。



平均数的标准误差是众多样本的平均数与其总体平均数距离的平均数。

2.8.3　比例估计

当调查一群人组成的样本，并且结果用某个百分比或比例来呈现时（比如，72%的水手患有膝部关节炎），那么这个百分比会与调查整个总体得出的实际百分比存在一定距离。如果这个样本是随机选取的，那么比例标准误差就表示样本百分比和总体百分比的接近程度。

比例的标准误差基于样本量和比例的大小。其计算公式如下：

$$\text{比例标准误差} = \sqrt{\frac{(\text{比例})(1-\text{比例})}{\text{样本大小}}}$$

和平均数的标准误差一样，随着样本量增加，比例的标准误差会降低。如果你有数学头脑，你也许会注意到，这个比例偏离0.50的程度越大，公式上半部分的数字就变得越小。

因此，当我们进行计算时，样本比例偏离0.50的程度越大，比例的标准误差就越低。另一个有趣之处是，公式的顶部是样本变异量的指示。(比例)(1−比例)是比例标准差的平方。



比例的标准误差是样本比例和总体真实比例之间距离的平均。

2.8.4　对未来表现的估计

在回归分析里，用一个变量或多个变量上的分数来估计另一个变量的分数[Hack #13]。但是，被预测的分数很可能不完全正确。

正如我们能够计算样本平均数和总体平均数之间距离的平均值，或者我们的调查结果和理论总体结果之间的距离，我们同样能够算出，平均来说，我们的回归预测结果和某个人实际获得分数的距离是多少。其计算公式如下：

$$\text{估计标准误差} = \text{标准差}\sqrt{1-\text{相关系数}^2}$$

方程式中用到的标准差是效标变量的标准差，效标变量就是你预测的变量。相关系数是你的预测变量和效标变量间的相关。



为了提高准确性（毕竟本条Hack的重点就在于此），我应该指出之前给出的标准误差的估计公式不完全正确。但是，它提供的结果和这个更为复杂而正确的公式几乎一样。

$$估计标准误差=标准差\sqrt{(1-r^2)\frac{样本大小-1}{样本大小-2}}$$

注意，这个公式里，相关系数越大，估计的标准误差就越小。这是合理的，因为如果两个变量间有很多信息重叠，你就能通过观察一个变量的分数，对另一变量的分数形成很好的概念。



估计的标准误差是实际分数和每个预测分数之间距离的平均值。

2

2.8.5 标准误差的运用

以下是如何使用这些方法，从而有一定把握来断定真相落在哪个区间。因为取样误差是正态分布的，标准误差可以和标准差一样，用来定义在正态曲线下分数的特定比例。

比如说，如果想要提供一个总体的值有95%落入其中的值范围，我们可以围绕我们的样本值建立95%的置信区间。基于正态曲线[Hack #23]，样本值左右1.96个标准误差应该能够提供一个范围值，我们能有95%的把握说这个范围值包含了总体的值。

表2-11展现了一些标准误差，以及使用样本数据来计算置信区间[Hack #6]的例子。注意一个更大的样本是如何创建一个更接近总体值的样本估计，同样，更大的样本量会指向一个更加准确的置信区间。

表2-11：建立95%的置信区间

标准误差类型	标准差	样本量	样本值	标准误差	95%置信区间
平均数标准误差	15	30	100	2.74	94.63~105.37
平均数标准误差	15	60	100	1.94	96.20~103.80
比例标准误差	0.25	30	0.50	0.09	0.32~0.68
比例标准误差	0.25	60	0.50	0.06	0.38~0.62
估计标准误差	15	30	100	14.81	70.97~129.03
估计标准误差	15	60	100	14.65	71.29~128.71



表2-11中估计标准误差所对应的“样本值”那一列，是对某个变量的估计或预测分数的例子。例子中的这两个计算假定预测变量和效标变量之间的相关系数是0.25。

2.8.6 弗兰克叔叔的捕狗员竞选

我叔叔弗兰克最近在竞选捕狗员职位，作为他的竞选经理，我有机会使用我掌握的标准误差

知识。在竞选的前几周，我在弗兰克叔叔居住的堪萨斯的Tonganoxie镇随机调查了30名投票人。我的调查显示50%的受访者表示会投我叔叔一票。我警告弗兰克叔叔，这个样本太小，不能非常准确地反映全体投票者的意愿。

在查阅表2-11后，我认为如果对全镇所有的投票人进行调查，他们把票投给弗兰克的百分比可能会合理地落在32%和68%之间，虽然最可能的值是50%。当然，我叔叔这个乐观主义者，将这解释为他可能有68%的选票，因此拥有巨大的领先优势。他将剩余的竞选专用款都花在了一场大型选前庆功宴上。作为一个现实主义者，而且深知我叔叔在小镇里的名声，我本人认为结果会朝相反的方向发展。结果的确如此。但那没什么关系。那是一场不错的宴会。

2.8.7　生效原理

如果我们遵循下面的假设并运用一些常识，我们就能信任标准误差的准确性。

- 取样误差是正态分布的

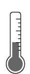
这意味着这些误差的大小以一种符合正态曲线的形式分布。这样我们就能够计算这些有足够说服力的准确的置信区间。

- 取样误差是无偏的

这意味着样本值大于或小于总体值的可能性相同。这样非常方便，因为这意味着通过重复的研究，我们可以逐渐接近真正的总体值。

公式以这样一种形式构建：如果你拥有少量总体信息或没有总体信息，样本估计中的误差大小约等于总体中标准差的大小。

看看样本量为1的时候，平均数的标准误差或是比例的标准误差会是多少，或者，当相关系数是0.00时，估计的标准误差会是多少。直观来看，一个好的计算标准误差大小的公式应该做到总体信息越多，产生的误差越小。



HACK #19

2.9　公正取样

如果你想了解企业的每位顾客或员工的情况，可以找他们每一个人谈话。如果你关注自家酒吧里出售的啤酒质量，可以在上酒前把每一杯都尝一尝。或者，为了节约时间、金钱和脑细胞，不妨代之以高效的“取样”。

健康的管理有赖于熟谙每个产品细节、每一笔交易和每一位客户的特点。当然，你永远无法将所有这些产品、交易和人整体都带到同一个显微镜下进行观察和评估，因为没有足够大的样本载玻片。

在社会科学领域也是一样，以人作为研究对象的学者不可能测量每一个人。就算我们有心尽可能多地刺探隐私，惊动别人，打断人家的事务，给人添麻烦，让人尴尬，换句话说就是打搅世界上的每个人，也不可能做到。我们没有足够的时间、空间和金钱，坦率地说，没人真的想要了解这么多人。

需要面对的问题是："如果不一一查看，又怎能了解每件事？"正如这本书介绍的所有Hack一样，统计学能提供解决办法。有很多科学合理的方法，让你通过观察任何事物的一小部分，就能准确描述其总体。

2

2.9.1 使用样本进行推论

推断性统计使我们能够基于小量的样本数据，引出一般性的总体结论。然而，要使这种推广有效，样本必须公正地代表总体。



总体，按照这里使用的意思来看，极少等同于社会学研究用语中"一个国家、城市或星球的全体居民"的意思。在推断性统计中，**总体**一词描述的是作为研究对象的某一类人或事物。比如，内布拉斯加州所有小学三年级的男生，堪萨斯州梅利亚姆市肖尼米什医疗中心（Shawnee Mission Medical Center）的护士，南美巨型水獭，或是美国国会图书馆的藏书。唯一的规则是总体要大于其对应的样本。

一个良好的样本能代表一个总体。这意味着总体中每个重要特征的分布必须和样本中这些特征的分布成比例。本项Hack大部分都是关于如何构建一个良好样本的，所以我们先来看一个好的样本。

想象一个由正方形、菱形和三角形构成的总体，如图2-4所示。

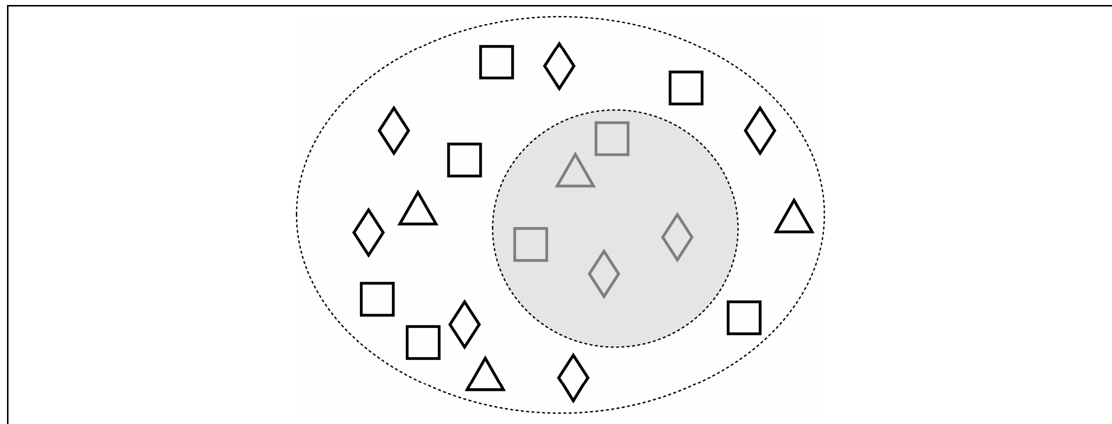


图2-4：总体中的一个样本

从正方形、菱形和三角形构成的总体中抽取一个公正的样本，其中包含这些形状，比例与它们在总体中的比例相等。在示意图中，外层的椭圆代表一个总体，其中不同形状的分布比例如下：正方形占40%，三角形占20%，菱形占40%。内层的椭圆代表样本，包含着总体中这些元素的一个子集。样本中各种形状的比例分布和与总体中各种形状的比例分布相同：40%的正方形，20%的三角形，还有40%的菱形。

这个样本是公正的。它很好地代表了总体，至少在形状特征上如此。当对人或对物取样时，样本通常呈现出多种多样的特质。人或物不完全是三角形或正方形，所以对于一个由人构成的样本来说，当其中某种特质的平均水平很好地匹配了总体水平，这个样本便具有代表性。在本例中，各种形状均为单一特质，而人则不然，各种特质在一个人身上可能或多或少地都有所呈现。（尽管根据我婶婶海洛薇兹的说法，我叔叔弗兰克是个“方正”[2]至极的人。）



提出问题的人必须选择他感兴趣的总体，其后他引出的结论只针对那个总体时才是正确的，而不适用于任何其他总体。

如果你知道构建该样本（内层椭圆中的元素）的采样方法是正确的，那么你可以仅通过观测样本来得出总体推论。其方法简单而又直观。

(1) 观测样本。比如，样本中三角形占20%。

(2) 对总体进行推论。我敢说三角形在总体中占20%。

我们且把理论总体中的抽象三角形放在一边，来看一个具体的例子。假设你想检查自己酒吧里出售的啤酒质量。为了解啤酒的总体质量，你需要构建一个良好的在售啤酒样本，然后逐一品尝样本。

(1) 观测样本。比如说，其中20%余味发馊。

(2) 对总体进行推论。我敢说你售卖的啤酒里余味发馊的占20%。你可能需要清理啤酒龙头。

推论很容易做，但只有当样本良好时，推论才是有效的。关键是构建一个良好的样本。

2.9.2　构建最好的随机样本

良好的样本代表了总体。代表性取样首先得定义好全集，即研究者想要从中取样的总体。在选择总体和选择样本时，对元素和各种隐性或显性的分组层级进行概念化的方法有很多种。你必须了解这些组织总体的方式，否则就无法创建好的样本。

注2：square除了指方形以外，还有为人诚实公正之意。——译者注

- 通用全集

指研究者希望将其结论推而广之的抽象总体。比如，我针对的可能是所有的漫画书收藏者。

- 可操作全集

可进行抽样的具体总体。比如，我不能完全确定自己已经找到或计算出了所有的漫画书收藏者，但我能通过将其定义成“所有的《漫画买家指南》（一本大多数认真的收藏者都会阅读的月刊）订阅者”，将总体变得可操作化。这种可操作总体不完全等同于通用全集，但它应该几乎和抽象总体一样大，而且能包纳研究者关注的抽象总体中的大多数。

- 抽样单元

指定义总体的元素。在本例中，每位杂志订阅者就是一个抽样单元。

- 抽样框架

指总体中抽样单元的列表，无论是真实的还是想象的。在本例中，抽样框架是杂志订阅者列表，我也许能够从杂志编辑部买到。



如果一个观察结果对于样本范围之外的人或事物可能有效，我们称之为**可泛化推广的**。如果一个样本不代表一个总体，那么这个样本就是**有偏样本**（一个坏样本）。

毫无疑问，最好的抽样策略是从有效的抽样框架中随机抽样。随机选择能够最好地创建一个能代表总体中所有被关注特质的样本。但是，随机选择的真正力量在于，抽样结果也代表了你根本没考虑到的、可能影响到观测结果的所有类型的变量。

从技术上讲，“随机”这个词描述了这样一个抽样过程：它给予总体中每一个成员相同和独立的入选机会。相同意味着抽样框架中的每个抽样单元和其他抽样单元拥有同样的机会。独立意味着一个人或一件事被选中的几率和其他特定的人或事是否被选中没有关联。

所以，假设有这样一个选择过程：按客户名单打电话，询问他们是否愿意参与活动，但如果第一次致电发现该客户不在家或者不在办公室，就放弃继续联系，这种做法没有给予所有可能的参与者相同的入选几率，不容易联系的人被选中的可能性较小。如果一个办公室里有人被选中时，就不再邀请同办公室的其他人，那么总体中每个成员的入选几率就不是独立的。

随机抽样可以通过这种方式来完成：用数字标记抽样框架列表里的所有名字，然后用某种随机数字选取法来选择每个被试。

2.9.3 现实世界的抽样策略

在现实世界里，随机抽样往往很难或者说是不可能的。下面是一些抽样策略，虽然不如随机

抽样，但在一些想象的科学实验室之外，却更加现实。

- 方便抽样

样本选择基于可得性。有时候也称作偶遇抽样。去本地购物中心，询问你最先遇到的10个人，了解他们对你公司产品的态度，这就是方便抽样。

- 系统抽样

单元是从抽样框架中等距抽取的。比如，你可能会从一个很长的人员列表中逢10抽取一个。只要列表中的人名顺序和你要判断的内容没有关联，这个方法对总体的代表性可能不亚于真正的随机选择。关于这个问题，统计学理论家和实践家之间实际上存在一些学术争论。

- 分层抽样

抽样框架被分成有意义的子群组，单元是从每个子群组里随机抽取的。如果定义子群组的特质对你提的问题很重要，那么这种方法可能会产生一个比随机抽样更具代表性的结果。

- 整群抽样

单元群组是随机选择的，这些群组中的单元都作为样本。例如，你可能会随机选择一家出版公司，然后就如何在出版界取得成功的话题访谈每一位员工。

- 判断抽样

其样本的选择是基于你的专业判断，决定这个样本能否代表总体。你也许会选择只和最佳客户们谈话，因为他们对你的产品最了解。

2.9.4　选择样本量

如果你能构建一个符合上述定义的良好的样本，那么即便是小样本也可以有效。不过，就像巧克力脆片曲奇饼的例子一样，我们的样本也是越大越好。样本量越大，就越能代表总体。因此，这样的观测结果更具有可泛化推广性，你也能更加确信其准确性。

同样，如果观测显示，变量之间存在某种有趣的相关，而且当你观测样本中的多个元素时，发现此种相关的可能性肯定比只观察少量元素时更大，你便能确信这种相关不是随机发生的。

最后，如果你的抽样的确出于某种社会科学的假设，那么在技术上必须符合特定的统计特征才能进行某种分析。在大样本中（譬如包含30个或更多部件的样本），这些标准更容易满足。

2.9.5　参阅

“看清实际错误程度”[Hack #18]介绍了如何在推断性统计中确定误差大小。



2.10　品尝苏格兰威士忌抽样

当统计学家从总体中选取人群样本时，他们实际上是从连续分布的变量中抽样。不过有些时候，当你把变量看作离散对象而不是连续分数时，更容易理解抽样的概念。

一些最强大的统计方法，是在等距测量或更高层次上[Hack #7]使用分数进行测量。但是，社会科学研究人员从总体中抽取分数时，总是选择人而不是分数。然后对人进行测量，产生一个分数的样本。迄今为止，这种做法的效果一直很好。

但是，说到抽样过程，精明的研究人员在取样策略上有时似乎并不那么精明。比如，如果一个研究者有意测量某个连续变量上的作用效果，比如*幸福感*，他也许会说（并且这么想）："好的，首先我需要一个样本，其中只包含幸福的人和不幸福的人。"至少在这个思考的瞬间，他是把幸福感作为一个二分变量来看待的。



二分是一个统计术语，表示"只有两个值"。比如，生理上的性别就是一个二分变量。

在他眼中，人们要么完全幸福，要么完全不幸福。当然，在现实生活中，他知道描述幸福感的分数范围区间是很广的，正因如此，他才用统计量做等距测量假设。

他把他的被试看做*非此即彼*（*即不是幸福就是不幸福*），是因为这么做能让他更容易描绘其抽样的代表性。这是一个聪明的策略，因为通过把样本视作大的分类变量的代表，而不是更精确的连续值，有时候能让抽样问题变得更容易回答和证明。

2.10.1　一个抽样问题

请看一个聚焦于抽样问题的难题。一个喝醉了的、无任期教职的统计学家（我见过不少）正在一个聚会上调酒。他在为他的系主任做威士忌苏打。主任要求威士忌和水达到某一精确比例(具体数字是多少并不重要，因为我们的主人公永远做不到那一步)。

这位统计学家首先找来两个容量相同的杯子。第一杯盛有2盎司苏格兰威士忌；第二杯盛有2盎司水。他开始从盛水的杯子里往盛有威士忌的杯子里倒了1盎司的水。显然，他已经搞砸了，因为他改变主意了，他把刚兑好的混合液体（3盎司威士忌和水的混合物）又倒回了1盎司到盛水的杯子里。现在两个杯子里都有2盎司的液体，但每杯的液体均为某种比例的水和威士忌的混合物。

这名统计学家很紧张，他试图重新开始，但是被系主任拦住了。系主任对他说：

“我有个提议：现在我们不可能知道每只杯子里威士忌和水的确切比例，因为我们不知道它们是怎么混合的。但是，如果你能正确回答下面的问题，我就为你向教职评定委员会写一封强有力的推荐信。如果你答错了，那么我可以肯定，凭你的资质，在酒店/汽车旅馆或是食品服务业找份工作应该不成问题。我的问题是这样的：现在，是第一杯里的水多一些，还是第二杯里的威士忌多一些？”

可以把这个问题想象为抽样问题。是第一个样本（即第一杯里的液体）中含有更多的水，还是第二个样本（即第二杯里的液体）中含有更多的威士忌？因为威士忌和水均由细小的粒子组成，很难想象每个样本代表的每种液体的量。即使按比例算，我们也不能确定有多少水粒子（或者说“水”的样本分数）混入了“威士忌”的样本分数，因为没人知道有多少水沉到第一杯的杯底部分，同时有多少留在顶部的酒被倒回了第二杯。这时人们需要凭直觉给出答案。令人遗憾的是，这个答案是错误的。

聪明人通常想到的直观答案是：第一个杯子（即开始盛有威士忌那杯）比开始盛水、后来又掺入威士忌的那一杯里的水更多。这似乎说得通，因为起初倒进威士忌里的是纯水，而后来被倒回水杯的则是水和威士忌的混合物。令人惊讶的是，这个聪明的想法把我们引入了迷途。正确答案是两个杯子里混合物的比例完全相等！威士忌杯中的水和水杯中的水含量相同。

2.10.2 使用比喻来解决问题

如果我们把此例中的变量想象成某种较大的物体，比如蓝色和白色的弹珠，而不是细微的粒子，那么问题的答案会显得更清楚。把一杯威士忌想象成一只装有100个蓝色弹珠的杯子。把一杯水想象成一只装有100个白色弹珠的杯子。

假设杯子很大，所以里面的弹珠能够很好地相混，就像混合液体一样。想想那种大玻璃鱼缸。这对确保选择的随机性很有必要。注意，睁大眼睛，在混合的每一步牢牢盯住这些弹珠。

我们的主人公从第二个杯子拿出50个白色弹珠，将它们混进第一个杯子。现在这两个变量的分布是：

- 样本1

100个蓝色弹珠，50个白色弹珠

- 样本2

50个白色弹珠

现在，他又从第一个杯子里随机取出50个弹珠（记住是随机的，以便模拟液体的混合），然后将它们混合到第二个杯子里。让我们想象一下各种可能的结果。

如果他碰巧选的全是白色弹珠，那么这些白色弹珠回到了第二个杯子，现在的分布情况是：

- 样本1

100个蓝色弹珠

- 样本2

100个白色弹珠

如果碰巧他连一个白色弹珠都没有选到，而是把50个蓝色弹珠放到了第二个杯子里，那么分布就是：

- 样本1

50个蓝色弹珠，50个白色弹珠。

- 样本2

50个白色弹珠，50个蓝色弹珠。

现在，想象一个更加可能的情景：他随机抽取的弹珠，一部分是白色的，一部分是蓝色的。比如，他可能抽出了10个白色弹珠和40个蓝色弹珠，然后把它们放入第二个杯子。在这种情况下，新的分布为：

- 样本1

60个蓝色弹珠，40个白色弹珠。

- 样本2

60个白色弹珠，40个蓝色弹珠。

按照这个方法，尝试你想要的任意一种弹珠混合方式，但是记住抽取的总数必须是50个（这是为了复制之前配酒的情境：往盛有威士忌的杯子里倒回1盎司水与酒的混合物，即杯中液体的一半）。

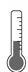
注意，无论你尝试哪一种混合方式，最后的结果都是每个杯子里各有100个弹珠。此外，最重要的是，注意最终第一个杯子里蓝色和白色弹珠的比例，始终等于第二个杯子里白色和蓝色弹珠的比例。任何不在第二个杯子里的蓝色弹珠必在第一个杯子里，任何不在第一个杯子里的白色弹珠必在第二个杯子里。

对于威士忌和水来说道理也是一样。正确的答案是它们的比例一定相同，不管最初是怎么混合的。

2.10.3　其他生效领域

现实生活中的民意调查公司要靠预测选举结果的准确性吃饭和维持自身名声，它们同样主要关心不同关键类别里每一类样本的比例。如果只有两个候选人，那么刚刚投完票的人不是把票投给了候选人A，就是投给了候选人B，即没有投给A的人必定投给了B。在一个类别里的缺失保证了在另一类别里的存在。以百分比的形式来报告预测结果带来了提高准确性的可能，但同样可能产生更大的误差：因为预测一个选民属于A类，结果却出现在B类，那么就在两个类别中都形成了误差。

当社会科学研究统计人员想要确定他们的样本能够代表总体时，他们主要关心的是特征在样本中所占的比例，而不是拥有这些特征的人数。最重要的是关键研究变量上，样本中每个分数的比例和总体中每个分数的比例相同。



HACK #21

2.11　选择可靠的均值

数据驱动的决策，比如判断自己在新城市是否买得起房，或者在生意上校准核心市场，总要依赖“均值”作为对大数据集的最好描述。问题是，有三种完全不同的值都可以被称作“均值”，而且它们往往导致不同的决策。所以，在决策中要注意选用正确的均值。

大多数人听到“这个镇里的平均房价是29万美元”（你可能觉得这一价格水平很便宜，也可能觉得很贵或者适中，这取决于你在哪个地方安家），他们会认为这个数字是通过加总镇上所有房屋的销售价，再除以房屋总数而算出的。但是统计学家知道，计算“均值”的方法不止一种，有时候其中一种比另外一种更好一些。

29万美元的价格是否真正代表了典型房价，取决于这个均值究竟是平均数、中位数还是众数。它同样取决于所有平均数据的分布形状。聪明人会确保决策中使用的是最佳汇总值。以下讨论的是每种均值的信任场合。

2.11.1　趋中趋势的度量

计算一组值的均值，无论它们具体是房价、期末考试分数，还是上瑜伽课的学生数量，目的都是为了有效传达这些值的*趋中趋势*。的确，大多数时候，趋中趋势是通过加总分布中的所有值，再除以这些值的数量之和确定的。然而统计学上并不把这个称作*均值*，而是称作*平均数*。那么，为什么不总是用平均数来计算趋中趋势呢？因为在一些情况下，平均数不能代表任何*真实值*！

请考虑本节开头提到的房屋均价的例子。假设你收集了镇上300所房屋的数据，想要计算这个样本中的平均售价。一般来说，平均数不能很好地指示房屋价格的趋中趋势。原因见图2-5。

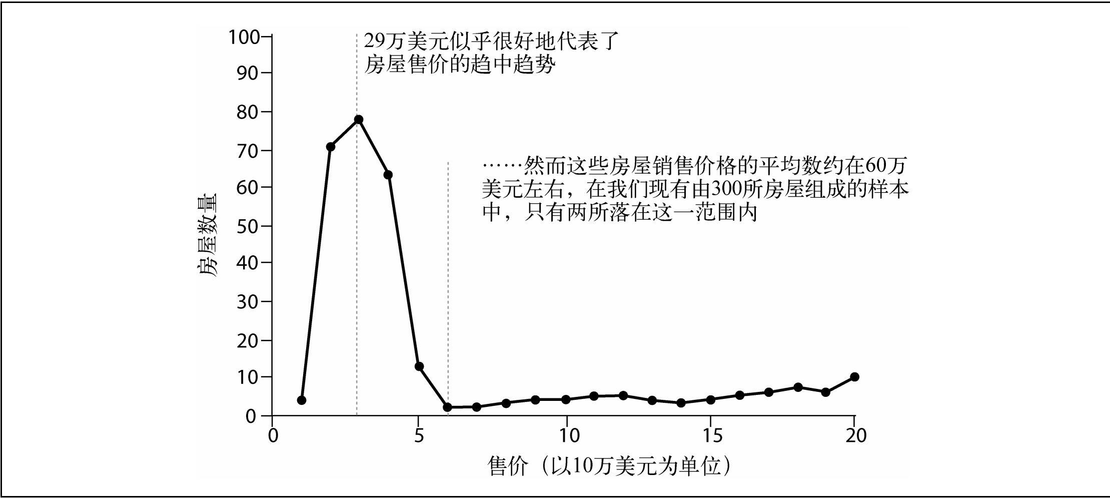


图2-5：平均数对均值的误导

在这种情况下，平均数不是非常可靠的均值，因为销售价格的分布被一些偏离中心的极大值所歪曲。在由300所房屋组成的样本里，231所房屋的售价介于10万美元到60万美元之间。剩下的69所房屋，售价在60万美元以上，其中56所超过了100万美元。平均数受到这些极大值的严重影响，因此无法代表样本中的任一房屋。

在大多数以金钱作为变量的情况下，都不太适合以平均数作为均值。依据平均数报告的平均收入指标总是高于大部分人的收入水平。因为总有少数像比尔·盖茨和J. K. 罗琳这样的人，会把平均数拉高。

那么，对这种类型的值，什么才是“有效的平均”？对于类似图2-5中的分布，可靠的统计学家倾向于报告中位数，而不是平均数。中位数是在分布中处于中间位置的值，即整个分布中有一半的值低于它，另外一半的值高于它（就好比高速公路中央的那条线，把路面分成两半）。在这个例子中，数据分布的中位数恰好是29万美元，因此它能很好地度量趋中趋势。

2.11.2　选择中间地带

中位数在这些情况下表现不错，因为与平均数相比，它对极端值的敏感性要低得多，因此当分布是正偏态分布或负偏态分布时，统计人员更倾向于采用中位数。故而，当分布被一些远远小于其他值的极端值所歪曲时（如图2-6所示，此例为包括50个学生测验分数的虚拟集合），中位数也被视作最“有效”的趋中趋势度量。

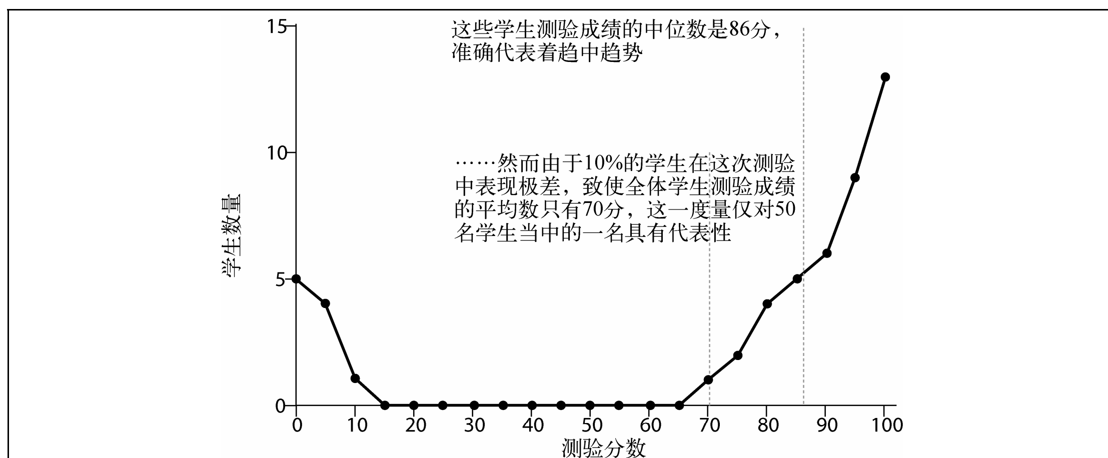


图2-6：中位数是对趋中趋势的最有效度量

图2-6显示了另外一种类型的数据，在这种情况下使用平均数可能导致错误的结论。以中位数作为度量，可以对班级分数得出更加准确的解释。

2.11.3　不适用领域

但是，即便是中位数也并不总是有效。考虑下面这种情景：假设你是一名瑜伽教练，你班里一半的学生年龄介于25岁至35岁之间，另外一半介于50岁至60岁之间。你会怎么描述学生的平均年龄？

像这种情况下的问题在于，无论平均数还是中位数都无法恰当描述这些个体构成的群组。那该怎么办？在这种情况下，最有效的均值选择是报告众数，也就是在数据样本中出现最多的值，如图2-7所示。

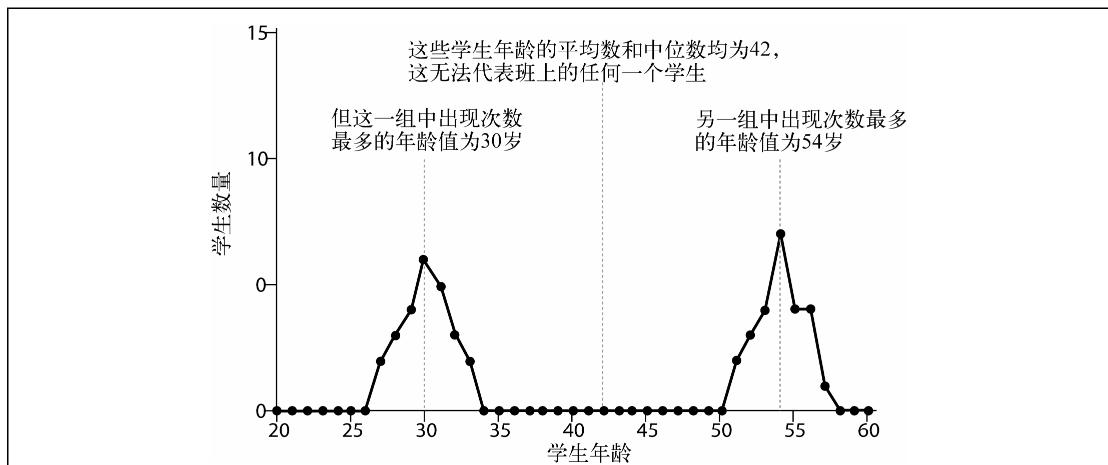


图2-7：作为最有效均值的众数

在这种情况下，有两个众数：一个是30岁，另外一个是54岁。一并报告这两个数值就是选择最有效均值的最佳方式。对于这种类型的数据，平均数和中位数都会形成误导。

2.11.4　如何选择有效均值

那么，什么时候平均数是有效均值？基本而言，当只有一个众数并且呈对称分布（意味着两端任意一个方向上都没有明显的歪曲）的时候，平均数是最佳的选择。如果你瑜伽班上的学员都是25岁到35岁，那么平均数便会是有效均值。

归根到底，应当如何选择最合适的均值呢？当你在报告汇总值的时候，遵循下面三条简单的原则，可以保证均值有效。如果你是基于这些数据做出决策的人，遵循这三条原则同样会让你作出有根据的选择。

- 如果数据中存在两个或更多的“趋势”（即两块或两块以上高频值区域），那么选择众数，并报告每个趋势的众数。
- 如果分布是偏态的（即一小部分极端值严重影响着平均数），那么选择中位数。
- 如果分布非常对称，且只有一个众数，那么选择平均数。

注意在大多数情况下，平均数、中位数和众数三者会非常接近，这很有趣。那么为什么要采用平均数呢？平均数始终是报告均值最常用的方式，因为如果我们想要获取另外一个样本数据并观察其趋中趋势，平均数更易于复制。中位数和众数的可变性较强，而平均数则比较友好和稳定。

——威廉·斯科朗普斯基



2.12　避开邪恶坐标轴

图形是表现数量、相关和研究结果的有力工具。但是，如果落入坏人手中，图形可能被用于欺骗目的。选择你的命运，年轻的卢克（或者年轻的阿纳金——如果你还不满25岁的话）[3]，切勿堕入黑暗面。

曾几何时，除了科学家、工程师和数学家以外没有人会关注图表。然而随着越来越多的新闻媒体瞄准大众市场，对数字信息的可视化呈现变得日益普遍。就拿昨天出版的《今日美国》（*USA Today*）杂志来说吧，其中至少包含了一打图表。

在商业会议上，也经常用图表来交流信息和论证所取得的成功（或失败）。如果创建图表时不够仔细，那么，一些看似随意的选择就会影响对信息的解释。你无需改变数据，就能改变数据的含义。

注3：天行者卢克和天行者阿纳金是美国科幻电影《星球大战》中的两位正面人物。——译者注

所以，在创建图表时，如果你想避免操纵受众，或者只想指出某个带有误导性的图表（不管这种误导有意还是无意的），那么不妨使用此项Hack来帮助你有效创建和解释图表。

2.12.1　选择可靠的图表

为了理解正确的和错误的绘图选项，首先需要介绍一些绘图的基础知识。图表中有各种各样的元素，通过操纵这些元素，可以正确地引导他人，也能造成误导。

典型的图表有两个坐标轴，因为它们描述了两个不同的变量。沿着底部的坐标轴称作*X*轴，而沿着侧边的轴叫做*Y*轴。



可以这么记：垂直的那个坐标轴叫做*Y*轴，因为这个可爱的小字母*Y*仿佛向上伸展着小手，垂直地指向天空。明白了吗？（欢迎来到充满创意的统计学教育领域。）

哪种图表适合（真实地）展示你测量的变量，取决于变量的测量标准[Hack #7]。你可以从三种常见的图表类型中做出选择，其中只有一种适用于你所测量的变量。

- 条形图

在图2-8中，*X*轴表示类别或组别，比如男性和女性。*Y*轴是连续变量：条形高度越高，变量*Y*的分值就越高。

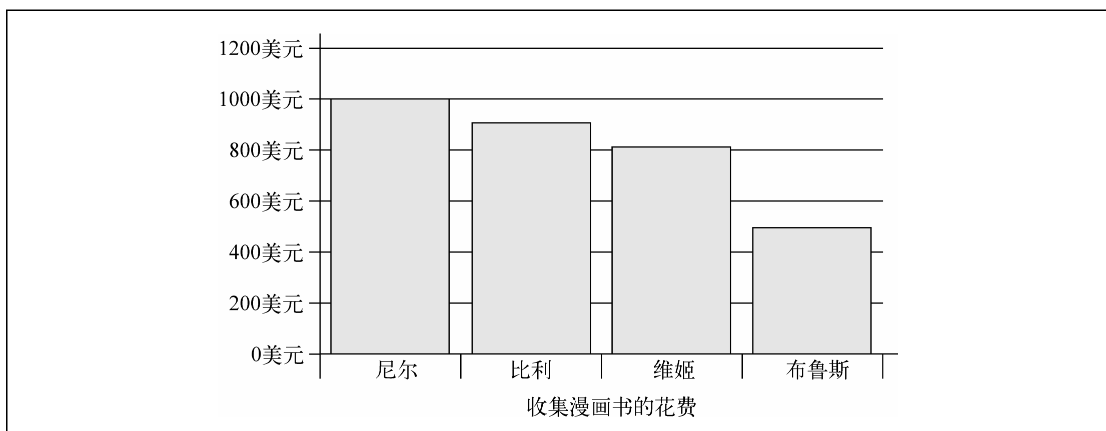


图2-8：条形图

- 柱状图

在图2-9中，*X*轴表示连续的值。柱状图总是运用于以下情况：*X*轴表示反映内在连续变量的普通类别，比如一年中的月份，或者其他可进行有意义排序的差异性分组。它和条形图看起来相似，只不过那些条形被挤到一起，相互间没有留下空间。

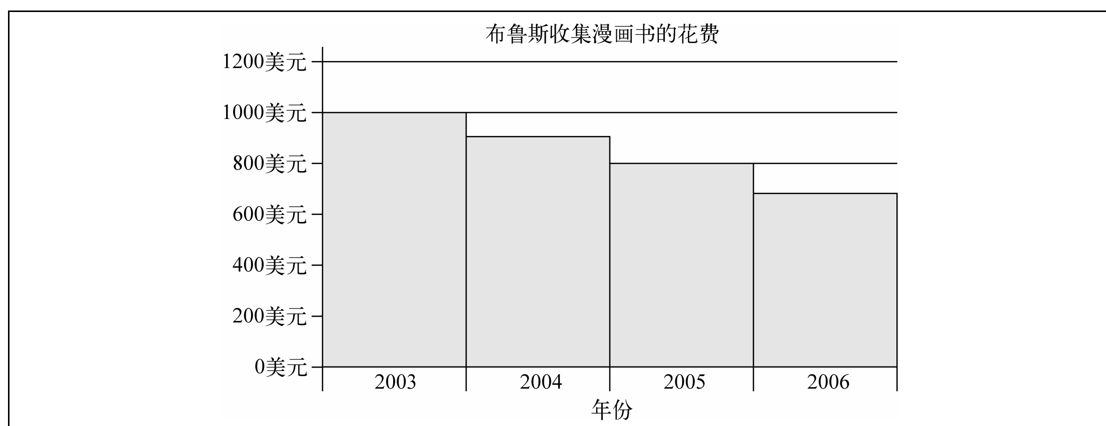


图2-9：柱状图

● 折线图

在图2-10中，X轴和Y轴都是连续变量，在这个例子中，它们分别表示时间和价值。线上的点位置越高，它在Y轴上的数量就越大。

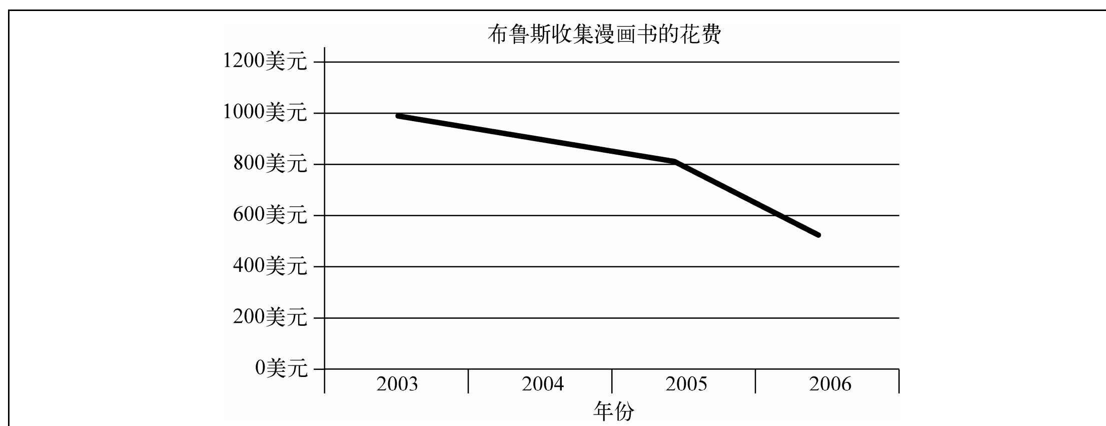


图2-10：折线图

为了选择正确的图表（即其样式具有最小欺骗性和最大直观性的图表），需要正确认定你在使用哪种类型的X变量（注意，Y在所有的样式中都是连续变量）。

- 如果X代表不同类别，而Y代表连续变量，使用条形图。
- 如果X可被看作分类变量，但是其次序仍有一定的意义，而Y是连续变量，使用柱状图。
- 如果X和Y都是连续变量，使用折线图。

2.12.2 图形暴力

绘图中的一个常见错误，通常与X轴的尺度设置有关，这也许是有意的，也许是无心之失。

接下来，让我告诉你这个问题的原因，以及如何避免它。

包含两个变量的图表引入了对比——它们或者是变量之间的对比，或者是时间跨度的对比，或者是一个变量上不同值的对比。一图胜千言，就像人们常说的那样，图表是非常具有说服力的证据。无论何时，当你使用折线图或是条形图来对比数值时，只有当线的高度或条形的长度是依据某个标准的最小值来得出判断，对比结果才是准确的。这一最小值通常是零。如果图表没有依据某些合理的基准值加以校准，那么实际上极其细微的差异在图中就会显得很大。

例如，对比图2-11中的两张图。它们所表示的数据完全一样，但是你对二者的解释可能差异极大。左上角的柱状图反映了美国股票市场在过去5天的表现。注意在第5天出现了一个看上去非常恐怖的下跌。毫无疑问，惊天动地的消息在第4天末就出现了。你也许同样注意到*Y*轴（道琼斯指数）的起点不是零，而是9900，一个低到足以包含所有5个条形顶部的值，但那样是没有意义的。

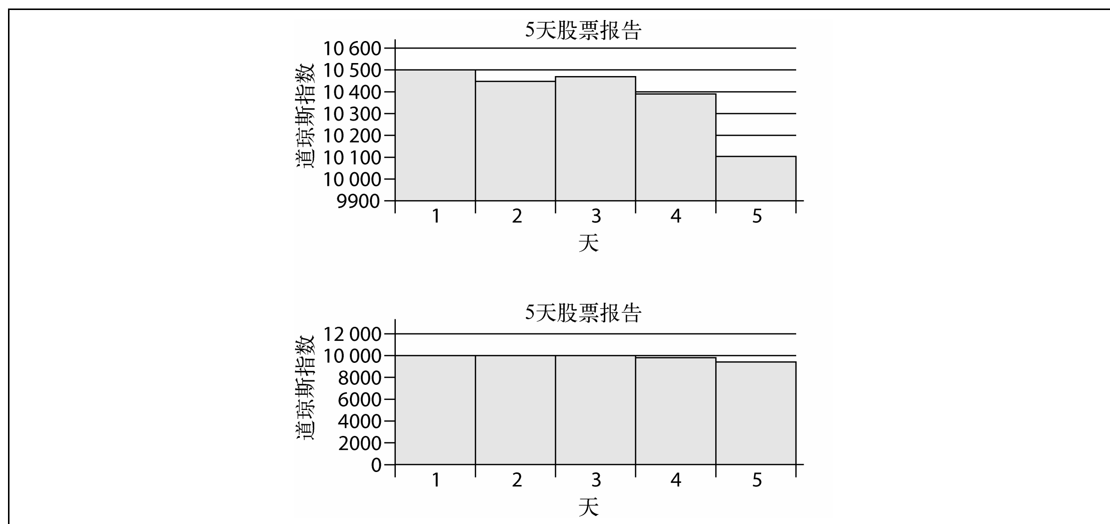


图2-11：*Y*轴的威力

让我们更仔细地看看图2-11中下方的第二张图。两张图展示的数据是一样的，但是第二张图的起点是0。这张图里展示的数据可解读为，股市在过去5天出现了小幅波动，第5天的可怕下跌只是暂时性的小问题。

这两张图哪个展示了正确的情形？二者都反映了从第4天到第5天股票市场有2.8%的下跌。究竟选择哪一种，实际上有赖于图表构建者的意图及其目标受众是谁。当涉及数字或是金钱时，通常并没有一个最有意义和最公平的起点。很多报纸提供的每日股票信息采用第一张柱状图的格式。他们认为读者对细微的变动感兴趣，所以将*Y*轴的起点值设置得尽可能高，以能够包含在*X*轴上的所有数据为底线。

但是，在一个经常改变投资组合，并且频繁买卖的贪婪的投资者看来，2.8%的下跌是很严重的事情。对于这一类读者而言，能够突出显示细微变动的图表设计也许最为有效。然而，如果一名投资者抱着“长期持有”的投资战略，那么相对微小的改变是没有意义的。

为了尽可能多地了解这类图表中包含的意义，需要经常检查*Y*轴的起始值。这样，当你查看图中那些条柱的时候，就能感觉到*X*轴上的真正差异。如果你正在绘制这样的图表，不妨想一想用哪种方式展示信息最为可靠。因为你的目的是如实传递信息，而不是欺骗——应该是这样吧？

2.12.3 参阅

《统计数字会撒谎》（*How to Lie With Statistics*，作者达莱尔·哈夫，1954年纽约，Norton and Company），该书首次向公众指出了图表的骗人花招，尤其是广告里的图表。

第3章

测量世界

（Hack #23~#34）

赋予现象以数量，对理解现象有很大的价值。虽然有时候将概念转换为数字的过程中，会丢失一些重要的信息，但是通过创建分数来代表任何我们感兴趣的事物，能够使理解更为精确，同样也使对比成为可能。这些Hack全都是关于分数测量以及分数解释的。

整个Hack家族都依赖于正态分布[Hack #23]，而正态分布无处不在。有了正态曲线，你能够知道和其他人相比你自己所处的位置[Hack #24]，能够在测试前预知可能的测试表现[Hack #25]，以及深度理解你的测试成绩[Hack #26和Hack #27]。

说到测试，你将学到如何编制一套好的问题[Hack #28]并进行一场高质量的测试[Hack #31和Hack #32]。你能辨别出糟糕的题项、无意义的问题，能够在不知道答案的情况下作出良好的测试表现[Hack #29]。你还能够在不研读书本的情况下，提高测试成绩[Hack #30]。

最后，通过学习一系列坚实的测量原理，你能够计算一个时代、一个人或一项事业的生命周期[Hack #33]，并学会使用可能延长寿命的医疗信息[Hack #34]。

测量又测量，本章内容全是关于测量的Hack。



3.1 看万物的形状

自然界几乎所有的事物都以相同的方式分布。只要你能够测量事物，不管它是什么，允许分数变化的情况下，它就有一个明确清楚的“正态分布”。如果你知道这种正态曲线的形状细节，就能对其表现作出非常准确的预测。

统计领域里有一些奇迹。至少有三种工具（或三种发现）是如此绝妙和神奇，以至于只要统计学的学生学到并开始理解它们的美，就会变得无比激动。

好吧，我可能有点夸大事实，但是的确有三种极好的理解世界的工具：

- 相关系数[Hack #11]；
- 中心极限定理[Hack #2]；
- 正态曲线。

因为我们已经在其他Hack中讨论了前两个奇迹的使用，现在我们把时间花在理解第三个（正态曲线）奇迹的形状和用法上。我很乐意展示这个能表现整个世界的正态曲线、正态分布、钟形曲线，如图3-1所示。

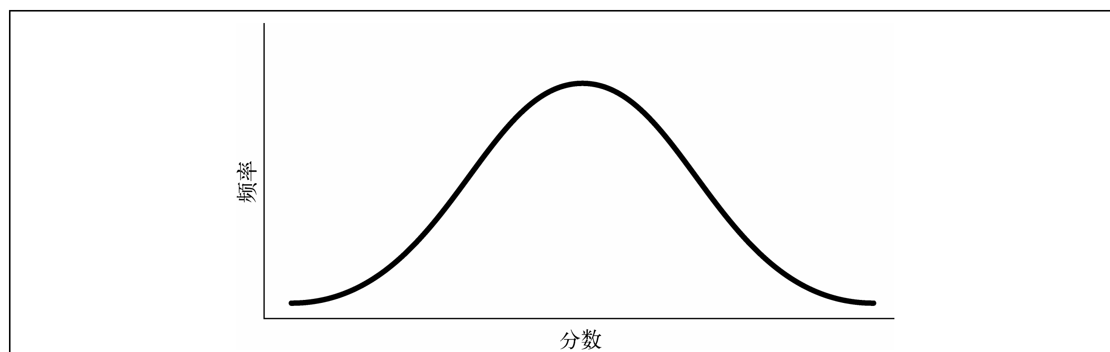


图3-1：正态曲线

3.1.1　应用正态曲线下的区域

统计学家已经非常详尽地定义了正态曲线。使用微积分和现实世界几百年的数据收集这两种方法可以发现，它们在关于正态分布的确切形状上得到的结论完全相同。图3-2展示了正态曲线的重要特征。平均数在中间，越偏离中心，分数的空间越小。

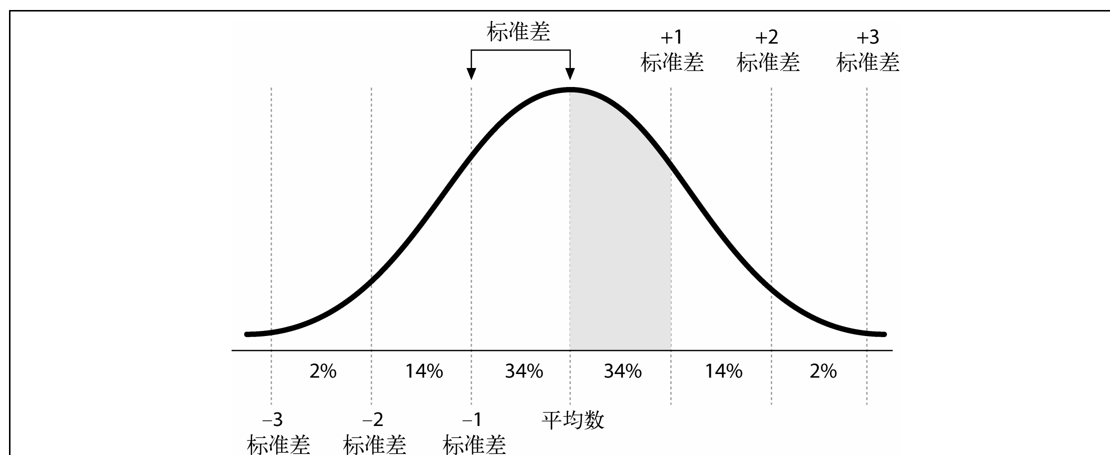


图3-2：正态曲线下的区域

虽然理论上正态曲线是无限宽的，但是平均数左右两侧的各三个标准差足以涵盖所有分数。



分布的**标准差**是每个分数离平均数的平均距离[Hack #2]。

1. 预测测试成绩

回想我之前作出的论断：你测量的任何事物都呈现为正态曲线。那么，言外之意就是，我们测量的任何事物的大部分分数靠近平均数，只有少部分分数远离平均数。测量足够多的人，你会偶然获得非常远离平均数的极端分数，但是这种分数非常罕见。获得特定分数的人群期望比例，随着分数远离平均数而变得越来越小。

那么你参加的下个测试会有怎样的成绩呢？我不知道有关测试或是有关你的任何信息，但我愿意打赌说你会获得一个接近平均数的分数。你也许会获得高于平均数的分数或低于平均数的分数，但是正态曲线告诉我，你的分数可能会非常接近平均数。

为了作出这类预测并对预测的准确性有十足的把握，你可以用已知的正态曲线来估计分数落入*X*轴上（图表的底部，水平部分）任意两点间的百分比。刻度上配对标准差点之间的分数百分比，如图3-2所示。百分比总和是100%，这是由于凑整导致的。记住，有些分数，虽然只有一小部分，但和平均数的距离超过三个标准差。

下面是有关曲线的几个重要事实，你能够利用这些事实去预测表现。

- 大约34%的分数落入平均数至平均数上方一个标准差内。看到图3-2中的阴影部分了吗？如果你拿一些墨水给正态曲线下方的整个区域上色，那你会在这块区域消耗掉34%的墨水。
- 大约34%的分数落入平均数至平均数下方一个标准差内。
- 大约14%的分数落入平均数上方的一个标准差至两个标准差范围内。
- 大约2%的分数落入平均数下方两个标准差至三个标准差范围内。

你同样可以组合这些百分比作出以下陈述，比如：

- 大约68%的分数在平均数正负一个标准差范围内；
- 大约50%的分数落入平均数下方。

你能用这些已知的百分比去做预测和概率解释。我们可以这么描述正态曲线：分数落入曲线的给定区域的百分比，也可以说任意给定的测试参与者落入给定区域的可能性。

- 下一次测试中，有2%的几率，你会得到高于平均数两个标准差的分数。
- 在我们的职业技能测试中，测试申请人仅有16%的几率会得到低于平均数一个标准差的分数。

2. 设定标准

政策制定者划分表现水平等级时，依赖于这样一个假设：能力是正态分布的。他们选择有这

样表现水平的人：有一定百分比能够胜任该工作。在制订录取政策或服务标准时，如果想奇迹般地提前知道会有多少人符合要求，正态分布是一个非常宝贵的工具。

比如，一所拥有高学术水准的大学，也许要求考生在一项能力测试中，分数至少高于平均数一个标准差。这样的话，他们就确保只招收能力在前16%的人。

同样地，美国特殊教育政策规定了学生在特殊教育状况资格测试（因此，是联邦政府和州拨款）上的分割分数（cut score）。分割分数是一个人必须高于（或低于）的特定分数。假设政策制定者只支付为2%的儿童设立的特殊教育项目和教员的预算，那他们会把分割分数设在平均数之下两个标准差处。对正态曲线的信任，使得他们能够计算出需要拨款的儿童数量。

3.1.2 体会正态曲线之美

为了体会正态分布的神奇，你可以经常建立自己的正态曲线。想象你测量了某件事物（比如态度、知识、身高或速度）。你有某个评分系统，分数允许变化（比如态度调查分数、SAT分数、英寸或每小时英里数）。因为你测量了很多人、很多建筑物或很多麻雀，所以你有很多的分数。现在，把这些分数放到一张图上，图中*X*轴表示从最低到最高的实际分数值，从左到右（或是从右到左，如果你喜欢的话）；*Y*轴（左侧垂直部分）代表分数群中每个值的相对频次。

在这样一个图表中，线或点的高度代表特定值分数的相对比例。注意：在正态曲线上，最高点位于中间，最低点位于两端。中间的分数是平均分，也是最大众的分数。在正态曲线上，中位数等于平均数，也等于众数[Hack #21]。

同样要注意，正态曲线是对称的：你可以将正态曲线对折，它的一边会完美地覆盖另一边。需要着重提及的正态曲线的另一个特征是：正态曲线是向两端无限延伸的。它是一条理论上的曲线，所以曲线的两端永远不会碰到底线。

正态曲线是联系自然万物的普遍真理。它是完美平衡的。它是永恒的。它是不朽的。它看起来有点像一只恐龙，非常酷。



3.2 计算百分位

理解测试成绩的一个简单而有效的方法是使用百分等级。下面讲解如何获取几乎没有解释价值的原始分数，并将其转换得更具信息量和价值。

在学校里，教师（或是顾问，或是任何报告标准测试结果的人）或许向你报告过结果，但从未报告过你的分数。取而代之的是，你可能看到一个看起来像百分比的数字，这个数字用来描述和其他参加测试的人相比，你（或你的孩子）的表现如何。这种类型的分数称作*百分等级*。

如果你得到了一个代表自己测试表现的百分等级，那么只有在你知道其含义的情况下，这个

百分等级才是有用的。另一方面，如果你必须向参加测试的人员解释其测试成绩，而你仅仅展示了一个原始分数，那么这种展示是没有意义的。建立或解释百分等级是一项对测试双方（参加测试的人和解释测试成绩的人）都非常有用的技能。

常模参照计分[Hack #26]是一种通过和其他分数进行对比，使测试分数更具信息量的一种方法。在现实世界里，你最常见到的常模参照分数是百分等级。百分等级被定义为"分布中低于给定分数的分数的百分比"。比如，在一项有20道题的小测试中，如果你答对了15道题，班里有一半的人答对的题数没你多，那你的百分等级就是50。

3.2.1　计算和报告百分等级

如果你是一名任课教师或人力资源经理，或任何必须向其他人报告测试结果的人，报告百分等级而非原始分数能够帮助参加测试的人员理解他们的表现，同样也能够帮助决策者理解设定不同表现标准的重要性。

1. 整理你的数据

计算百分位首先要整理你所有的测试分数。对于小数据集合，建立一个频次表非常简单，这个频次表除了能提供百分等级，还能回答各种问题。

下面是一个课堂测试中30个分数的样本分布（由最低到最高排列），100分是最高分：

59、65、72、75、75、75、80、83、83、85、85、85、85、85、85、86、86、86、86、88、88、88、90、90、90、90、90、92、94、97

2. 计算频次和百分位

为高效起见，可按表3-1展示这些数据，并计算每个分数的频次。

表3-1：课堂测试累计频次表

分数	频次	累计频次	百分比	累计百分比
59	1	1	3.33%	3.33%
65	1	2	3.33%	6.67%
72	1	3	3.33%	10.00%
75	3	6	10.00%	20.00%
80	1	7	3.33%	23.33%
83	2	9	6.67%	30.00%
85	6	15	20.00%	50.00%
86	4	19	13.33%	63.33%
88	3	22	10.00%	73.33%
90	5	27	16.67%	90.00%
92	1	28	3.33%	93.33%
94	1	29	3.33%	96.67%
97	1	30	3.33%	100.00%

表3-1展示了参加测试的人获得了哪些分数，有多少人获得了那个分数，获得给定分数的总人数，获得分数低于给定分数的总人数，获得某个分数的人数占所有人数的百分比，获得不高于给定分数的人数的总百分比。累计频次列总是报告出分布中的（在我们例子中是30人）总人数（或分数），以及人数总的百分比（总是100%）。

3. 计算百分等级

为了计算分布中任意分数的百分等级，需要使用“累计百分比”一列。找到感兴趣的分数，查看其所在行的上一行对应的累计百分比。比如，对于94分来说，百分等级是93.33，大约是第93个百分位。86分的百分等级是50。



如果你查看一系列统计学或测量学教材，会发现，对于百分等级来说，实际上存在两种不同的、有争议的定义。我更喜欢“分布中小于感兴趣的给定分数的分数百分比”这个定义，但是有些书给出的定义是“分布中**等于**或小于感兴趣的给定分数的分数百分比”。两个定义都是合理的，且在这两种定义下都可以通过频次表来计算百分等级。在第一个定义下，不存在第100个百分位。在第二个定义下，不存在第0个百分位。选择并使用你偏爱的定义，但记住，在呈现结果时要和大家分享你的定义。

3.2.2 解释百分等级

想象一下你对面坐着你的指导顾问，你被告知自己的百分等级是93。那么，这代表什么意思？好吧，最直接的解释是：在所有参加测试的人中，有93%的人得到的分数比你低。这么说同样也是正确的：有7%的人分数和你一样或高于你。我们同样能够把百分等级看做分数偏离常态的距离。平均百分等级总是在第50个百分位附近，如果分数是正态分布，那么平均百分等级正好就是第50个百分位。所以，我们同样可以说第93个百分位远远高于平均百分等级。

不要犯很多精明的统计黑客有时也会犯的错误。本Hack前半部分，我们使用了一个测试分数的例子，你在一项有20道题的小测试上答对了15道题，班里有一半的人答对的题数没你多。在这个例子中，你的百分等级是50。注意，此例中，你答题的正确率是75%（15/20），但是百分等级是50。不要把这两个概念搞混了！你的百分等级无法说明你究竟答对了多少道题。

3.2.3 不适用领域

记住，只有在你寻求常模参照解释时，百分等级才是有用的。如果你想知道自己是否掌握了一系列关键技能，那么知道有多少百分比的人已经掌握了多于或少于这些的技能是毫无帮助的。为了知道和某套标准相比你所处的位置，而不是和其他人相比你所处的位置，你需要一个标准参照分数[Hack #26]。在这种情况下，正确率这一类型的分数比百分等级更有意义。

3.2.4 参阅

如果假定你的分数是正态分布的，或者说至少来自于正态分布的总体，你刚好能利用正态曲线下方区域的信息将标准分数直接转换成百分等级[Hack #25]。



3.3 利用正态曲线预测未来

在自然界中，我们测量的几乎所有事物都有一个已知的分布形状，即“正态曲线”，所以我们能够利用这个分布的精确细节来预测未来，并回答各种概率问题。

本书中，很多Hack #都充分利用了统计学家和正态曲线的密切关系。“看万物的形状”[Hack #23]展示了使用正态曲线预测测试表现的大体方法。但是，我们能够做得更好。

我们掌握了如此多的关于这条神奇曲线准确形状的信息，以至于能对分数落在某个范围内的概率作出准确预测。可以提出很多和测试表现相关的其他类型的问题，统计学能在我们参加测试前就帮助我们解答！比如：

- 你的分数落在任意给定两个分数之间的几率是多少？
- 有多少人的得分介于这两个分数之间？
- 你通过下次测试的几率是多少？
- 你会被哈佛大学录取吗？
- 在美国有多少百分比的学生能够成为国家优秀奖学金获得者（National Merit Scholar）？
- 我叔叔弗兰克通过门撒资格测试（Mensa qualifying exam）的几率是多少？

回答这种类型的问题，需要一个精确的工具。本Hack提供了所需的工具：正态曲线下方区域的表格。

3.3.1 正态曲线下方区域的表格

正态曲线由分布的平均数和标准差来定义，不管我们测量什么，只要计分系统容许分数产生变化，那么曲线的形状就总是相同的。落入曲线下方不同区域的分数所占比例已经被明确规定好了，比如不同标准差之间的空间以及距平均数的距离。

这个Hack依赖于一张看起来有些复杂的表格，但这张表格富含如此多的有用信息，以至于它会很快成为你黑客工具箱中的一个主要的工具。事不宜迟，让我们深呼吸，来看看表3-2。

表3-2：正态曲线的下方区域

z分数	平均数和z分数之间分数的比例	大区域中分数的比例	小区域中分数的比例
0.00	0.00	0.50	0.50
0.12	0.05	0.55	0.45
0.25	0.10	0.60	0.40
0.39	0.15	0.65	0.35
0.52	0.20	0.70	0.30
0.67	0.25	0.75	0.25
0.84	0.30	0.80	0.20
1.04	0.35	0.85	0.15
1.28	0.40	0.90	0.10
1.65	0.45	0.95	0.05
1.96	0.475	0.975	0.025
4.00	0.50	1.00	0.00

3.3.2　解密表格

在使用这个极好的工具前，我们需要再次深呼吸，然后了解一下情况。我已经用好几种方式简化了这张表的信息。首先，我只列举了一些能计算出数值的信息，并没有列出全部。事实上，很多统计学书以0.01为增长速率，列出了0.00~4.00的*z*分表数。那样会展示很多信息，此处我们截取最常用的一部分信息，包括达到90%置信区间（1.65）所需的*z*分数，以及95%置信区间（1.96）的*z*分数。想知道更多关于置信区间的信息，可参考“精确测量”[Hack #6]。

我把比例四舍五入至小数点后两位。最后，我在表格中用z符号以标准差的形式表示和平均数之间的距离。你能在“给原始分数改头换面”[Hack #26]中，学到更多有关*z*分数的知识。

理解了对表格所做的简化后，可以使用它对表现进行概率预测或回答统计问题，第一步就是理解第4列的含义。

- *z*列

描绘正态曲线[Hack #23]。你可能对某个可能落入底部水平线的分数感兴趣，而它与平均数也有一定距离。它可能比平均数大也可能比平均数小。用标准差表示与平均数的距离就是*z*分数。*z*分数为1.04，描述的是距离平均数1.04个标准差的分数。因为正态曲线是对称的，故而我们不用在意距离的正负，所以展示出来的*z*分数都是正值。

- 平均数和*z*分数之间分数的比例

在平均数和一个给定分数的空间内，存在某个比例的分数。这是一个随机分数落入由平均数和任意*z*分数所限定区域的概率。

- 大区域中分数的比例

你同样能够描述任意给定*z*分数和*z*分数为4.00之间的区域，或者说是曲线的末端。

理论上，曲线不会真正终止，但*z*分数为4.00已经非常接近涵盖100%的分数。

但是，曲线有两个末端。除非*z*分数为0.0，否则*z*分数和曲线一端的距离一定大于*z*分数和曲线另一端的距离。这一列指的是*z*分数和曲线最远端的区域，这一列的值是落入这个区域分数的比例。换句话说，是一个随机个体会在这个区域获得分数的几率。

- 小区域中分数的比例

这列指的是*z*分数和曲线最近端的区域。它表示落入这个区域分数的比例。

3.3.3 估计得分高于或低于任意分数的几率

如果你想知道被大学录取的几率，就要明确你需达到的分数，这个分数在学校入学测试中也被称作分割分数（cut score）。只要你知道了这个分数，就能找出这个测试的平均数和标准差。（所有这些信息可能都在网上。）将你的原始分数转换成*z*分数[Hack #26]，然后在表3-2中找到那个*z*分数，或是接近*z*分数的分数。

判断分割分数是否高于平均数。

- 如果分割分数高于平均数，查看“小区域中分数的比例”列。那代表你获得等于或高于分割分数的几率，以及你被录取的几率。
- 如果分割分数低于平均分（这不太可能，只是为了完整地训练你如何使用这个工具），查看“大区域中分数的比例”列。那代表被录取的学生比例，若其他因素等同，也代表你被录取的几率。

确定得分低于一个给定分数的几率时，步骤和上述提到的选择相反。分割分数低于平均数，获得低于特定分割分数的几率在“小区域”列。分割分数高于平均数，则获得低于给定分割分数的几率在“大区域”列。

3.3.4 估计得分介于任意两个分数之间的几率

要想知道获得一个介于某个计分分数（scoringscore）范围内的分数的几率，可以通过查看正常落入那个范围的分数比例来计算。

如果你想要知道有多少比例的分数落入曲线下方任意两个点的分数之间，那么通过*z*分数来定义这些点，并计算相关比例。根据两个分数是否落在平均数的同一侧，可利用下述方法计算介于这些点之间的分数的正确比例。

- 如果*z*分数在曲线的同一侧，查看“大区域”列或查看“小区域”列，得到两个*z*分数，然后用高值减去底值。
- 如果一个*z*分数落在平均数左侧，另一个*z*分数落在平均数右侧，平均数在这两个*z*分数中间，那么使用“平均数和*z*分数之间分数的比例”列。查看两个*z*分数值，然后将它们相加。

3.3.5 计算百分等级

这张表格的第三种用途是计算百分等级。你可以在“计算百分位”[Hack #24]中读到更多关于常模参照的内容。对于高于平均数的分数，百分等级是“平均数和*z*分数之间分数的比例”加上0.50。对于低于平均数的分数，百分等级是“小区域中的分数比例”。

3.3.6 判断统计显著性

这种表格的另一种用途是确定分数差异的统计显著性[Hack #4]。通过确定落入分数之间某个距离或更远距离的分数比例，你能计算出那个结果的统计概率。

更有用的是，其他的统计值，比如相关系数和比例也能被转换成*z*分数，这张表同样可以用来将这些值和0对比，或者进行相互对比。

3.3.7 生效原理

“看万物的形状”[Hack #23]提供了对正态曲线的很好概览。但是，仅通过在表3-2中查看这些值的改变方式，你就能感觉到正态分布的形状。平均数附近，每行有着较小的*z*分数，但有很大比例的分数落入。随着向远离平均数的方向移动，若要包含相同比例的分数就需要越来越大的曲线区域。

比如，*z*分数从1.65猛增到4，只覆盖了分布的后5%。但是，在平均数附近，*z*分数只需从0.12增加到0.25，就能覆盖分数的5%。这张表格证明了常见的有多常见，罕见的有多罕见。

3.3.8 参阅

你可以利用如下网址来计算自己的正态曲线下的准确区域：http://www.psychstat.missouristate.edu/introbook/sbk11m.htm。这个网站由大卫·斯托克伯格（David Stockburger）维护，里面有很好的讨论和交互式的计算器。当你访问此网站时，不要被Mu和Sigma这两个词弄糊涂了，它们是分别代表平均数和标准差的统计术语。



3.4　给原始分数改头换面

测试的原始分数意义不大甚至没有意义。但是，将可怜的原始分数转化成"z分数"后，你几乎难以相信有多少信息被塞进了这个小小的超级数字里。

那个直接获得且显而易见的原始分数（比如高中测试分数），传达的信息量非常少，这令人震惊。比如下面的例子。如果我从学校回家告诉妈妈我今天在学校的一项重要测试中得了16分，她可能会说些什么，比如"你42岁了，为什么还住在我家里？"以及"不错，亲爱的。16分算好的吗？"

当你只是告诉某人一个原始分数时，被分享的真实信息非常少。你不知道16是否算一个不错的成绩。你也不知道16是相对高还是相对低的分数。有很多人得到16分甚至更高的分数吗？还是很多人获得了低于16的分数？即便我们知道测试分数的分布范围和所有可能的分数等信息，也依然无法将这次测试的分数表现和过去测试或下次测试的分数表现进行对比，也不能和其他学科的分数对比。原始分数实际上是没有意义的。

不要烦恼！你依然能够知道你以及其他人的表现。你依然能够作出选择，并透过人和测试进行表现比较。依然有希望！

原始分数可以被转变成一个能做所有事情的新分数，那是97磅这种无能的原始分数永远都做不了的。原始分数能被转换成一个超级数字：*z*分数。不像原始分数，*z*分数会告诉你，你的表现是高于还是低于平均水平，并且会告诉你高于或低于平均水平的程度。*z*分数还能使你进行不同测试和事件的对比，甚至对比不同的人。

3.4.1　计算*z*分数

可以通过一种方式将原始分数转换为*z*分数，那么新的数字表示原始分数高于或低于平均数的程度。

下面是公式：

$$z = \frac{\text{原始分数} - \text{平均数}}{\text{标准差}}$$

为了将原始分数转换成*z*分数，先用原始分数减去平均数，然后除以标准差。分布的标准差是每个分数和平均数距离的平均[Hack #2]。

3.4.2　理解表现

*z*分数的值通常介于–3至3之间。仔细检查方程式顶部，你也许会注意到以下内容：

- ❑ 如果原始分数比平均数大，那么*z*分数为正；

- 如果原始分数比平均数小，那么*z*分数为负；
- 如果原始分数正好等于平均数，那么*z*分数为0。



*z*分数的值往往介于–3至3之间，因为分数的**正态分布**通常刚好是6个标准差的宽度[Hack #23]。

明智的测量专业人员在报告结果时会使用*z*分数技巧。你看到的全都是基于*z*分数的分数，通常称为标准分数[Hack #27]，而不是原始分数。这些标准分数有已知的稳定特征。因此，如果你知道这些分数的特征（它们的平均数和标准差），就能将它们转换回*z*分数，知道和其他人相比你的表现如何。

为了说明如何使用这个法则来揭示有关你表现的隐藏信息，我们以ACT测试为例。美国很多高中生都参加美国大学入学考试（The American College Test），很多大学也将其作为录取条件。ACT是一项成就和能力测试，被认为能够预测学生在大学的表现。

测试的每部分的分数范围都是1至36。虽然在过去几十年里，实际的测试描述性统计发生了波动（因为分数有提高），但官方报告的ACT平均数总是为18，标准差为6。想象3个学生参加ACT测试，得到3个不同的分数。我们可以用ACT分数分布的平均数和标准差将这3个分数转换成*z*分数，如表3-3所示。

表3-3：将原始分数转换成*z*分数

学生	ACT分数	(原始分数–平均数)/标准差	*z*分数
扎克	14	(14–18)/6=–4/6	–0.67
泰勒	18	(18–18)/6=0/6	0.00
艾萨克	24	(24–18)/6=6/6	1.00

扎克的*z*分数是负的，所以我们知道他的得分低于平均水平。他的得分低于平均数大概2/3个标准差。泰勒的*z*分数是0.00，表示和过去这些年参加ACT的其他人相比，他的表现处于平均水平。艾萨克做得最好，得到高于平均数1个标准差的分数。



每年举行测试的时候，实际的ACT平均数和标准差都会改变。过去几年真正的平均数和标准差大约是21和4.5。

3.4.3 确认你表现的稀有性

虽然知道和其他参加测试的人相比你的得分情况，比只知道一个原始分数更有用，但*z*分数真正的解释力来自于它和正态曲线的关系。图3-3是一张正态分布图，和“看万物的形状”[Hack #23]里展示的图相似。

“看万物的形状” [Hack #23]里展示的图和这个图的差异在于：图3-3将这些值作为*z*分数展示，而不是展示每个标准差离平均数的距离。通过使用正态曲线下区域的知识，你甚至能从*z*分数中学到更多的知识。如果分数是正态分布的，那么你可以就分数在某个区间出现的概率这一话题侃侃而谈了。

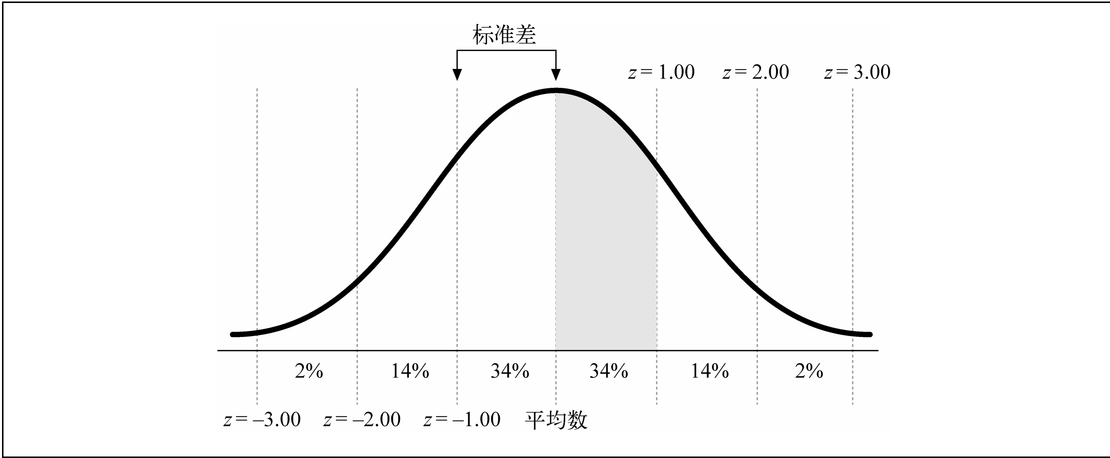


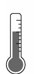
图3-3：*z*分数和正态曲线

表3-3中展示的分数同样能被解释为相比这名学生，表现更好或更差的学生人数。泰勒的*z*分数为0.00，这意味着他的表现比50%的学生要好。同样能够将学生的分数表述成概率。如泰勒有50%的几率会得到0.00或更好的*z*分数。在任意测试中，学生只有16%的几率能得到1.00或更高的*z*分数，所以相比其他参加测试的人来说，艾萨克做得很好。

3.4.4　生效原理

将原始分数转换成*z*分数后，我们就能和其他人进行对比了，你可能觉得这是合理的，其实不止你一个人这样认为。在过去的关于教育测量领域的100年里，社会学家（以及任何必须评估人类表现的人）一直被常模参照（norm-referenced）解释的简洁性吸引。即使不确定测试分数的真正含义，我们至少能将你的分数和其他人的情况做对比。不管我们测量的是什么，我们至少能够知道你所拥有的比其他人更多还是更少。

另一种用来解释教育和心理分数的方法是标准参照（criterion-referenced）。这种方法需要知道更多的有关我们已测量的特质或内容的信息，并且在事前就要确定需要多少信息量。标准参照测量使得每个人都能获得同样的分数，只要他们满足同样的标准。前一种方法（常模参照）以前是并且一直是最受欢迎的解释方法，而后一种方法（标准参照）才刚刚起步。


HACK #27

3.5 标准分数

令人惊讶的是，所有这些知名的、利益攸关的考试，比如SAT、ACT以及智商测试，从来都不报告你的原始分数。相反，测试报告已经将那个无用的数字转换成了一个更有意义的分数，这个分数可用来和其他参加同样测试的人进行对比，从而知晓你的表现如何。一旦你理解了“标准”分，就能自己计算标准分，甚至创建自己的标准分。

“给原始分数改头换面”[Hack #26]讨论了*z*分数的超能力。这些标准分数给无意义的原始分数添加了各种各样的信息。有一点非常好，那就是任何使用这本书的人都能解释*z*分数，并基于分数的解释信息作出决策。

但是，当你想解读很多分数报告（比如你刚获得的SAT分数）时，却发现在任何地方都不会看到*z*分数，相反，你看到的是一些奇怪的自定义标准分。这种自定义的标准分只有报告分数的相应公司才使用，它有点像*z*分数但是和*z*分数有差异，这种差异足以让新手感到分数是无意义的。

不要害怕。你可以利用下面的工具解释这些奇怪的标准分数，甚至创建自己的标准分数（当你向其他人报告你那奇怪的测试，而这项测试席卷全国时，你会像ACT先生或IQ小姐以及任何依靠测试赚钱的人一样富有）。

3.5.1 *z*分数的问题

可以确定的是，当向参加测试的人或他们的父母、大学和正考虑使用*z*分数的雇主汇报时，*z*分数存在某种缺陷阻碍了它的广泛使用。但另一方面，大多数测试公司在创建一个更有吸引力的标准分时，首先使用的是*z*分数，随后才报告更有吸引力的标准分。

使用下面这个公式将原始分数转换成*z*分数：

$$z = \frac{\text{原始分数} - \text{平均数}}{\text{标准差}}$$

正如“给原始分数改头换面”[Hack #26]里详细描述的一样，这个公式计算出的*z*分数往往在−3.00~+3.00，平均数是0.00，标准差是1。虽然*z*分数作为解释测试分数的工具非常有用，但人们看到它时，仍会由于一些原因不喜欢这些数字。

- 它可以是负值。事实上，有一半的*z*分数是负值。你很难说服参加测试的人相信负分不是坏消息。
- 0.00分是平均分！如果我无法向人们解释负分不一定是件坏事，想象一下，尝试说服父母说我们期望小比利在那个重大测试中得到0分，当他的确得了0分时我们很高兴（将会多么困难）。

- ❑ 你能预期的最高分是3.00，并且在100个参加测试的人里只有1个人能得到那个分数。如此一来，即使在测试准备时付出相当大的努力，看起来也只是为了获得微不足道的3分！

测量人员已经探索和发现了报告测试分数的其他标准尺度，这些标准尺度有着更令人满意的属性。诀窍是从计算z分数开始，然后将其转换到某个其他更友好的平均数和标准差的尺度上。

3.5.2　创建和解释*T*分数

z分数的一个问题是：平均数是0。把0分当做一件好事来报告，会使得一些教师、父母和学生因误解而不高兴。我们可以通过在字母表里把z移动到T的方法解决这个问题。

T分数是对z分数的转换。转换后新分布的平均数是50，标准差是10。T分数方程式使用了后向转换方法。下面是T分数的计算公式：

$$T=z\ (10)+50$$

所以，如果小比利在一项重大测试上的得分是平均分，那他得到了0.00的z分数，不要给他的父母报告这个吓人的分数，而是把它转换成T分数：

$$T=0.00\ (10)+50 \qquad T=0.00+50 \qquad T=50$$

报告比利得到了50分。恭喜！为了使分数更有意义，一名好的教师或者学校顾问会解释说T分数得范围是20至80，其中50分是平均分。

在一些测试报告中，相对z分数来说，T分数是个更好的备选方案。T分数不会是负值，平均数也是看起来比较可观的50分。



明尼苏达多项人格问卷Ⅱ（Multiphase Personality Inventory-II）是一项非常流行的、使用T分数分布报告分数结果的心理学测试，用以测量抑郁、精神分裂等。每个MMPI-Ⅱ子量表的平均分都是50，标准差是10。通过把每个子量表的分数放到同一个尺度上，你能够进行特质间的横向对比，并能建立一个分数概况（profile），从而更全面地了解参试人员。

3.5.3　创建自定义的标准分数

测试开发人员已经找到了其他报告标准分的方式。表3-4列出了很多知名的、利益攸关的测试，很多人参加过或将会参加这些测试。

表3-4：常见的标准分数分布

测试名称	典型分数范围	平均数	标准差
z分数	−3.00~3.00	0	1
T分数	20~80	50	10
美国大学测试（ACT）	1~36	18	6
SAT	200~800	500	100
美国研究生入学考试（GRE）	200~800	500	100
研究生管理科入学考试（GMAT）	200~800	500	100
法学院入学考试（LAST）	120~180	150	10
医学院入学考试（MCAT）	1~15	8	2.5
韦氏智力量表（IQ测试）	55~145	100	15
斯坦福比纳智力测试（IQ测试）	52~148	100	16

因为测试分数是正态分布的，所以你可以把任意分数放在正态曲线上来解读，从而看出你的表现是否处于平均位置，是否出奇地低或高[Hack #23]。

3

3.5.4 创建自己的标准分

为了好玩，你可以按照自己的意愿，以任意平均数和标准差来创建自己的标准分。难道你不希望自己的SAT分数是350分？选择一个分布范围，然后进行分数转换吧。

比如说，你偏爱这样一个分布：平均数是752 365，标准差是216 456。（谁不会？）我们把这个分布称作Frey分数分布。套用T分数规则，你能够把350分的SAT分数转换成Frey分数。记住，你必须首先将350分的SAT分数转化为z分数：

$$z=\frac{\text{原始分数}-\text{平均数}}{\text{标注差}}=\frac{350-500}{100}=\frac{-150}{100}=-1.50$$

然后将它转换成Frey分数。

$$\text{Frey}=-1.50\times 216\,456+752\,365=-324\,684+752\,365=427\,681$$

现在，427 681分是不是比350分听起来要好？因为你知道Frey分布的平均数，所以对两个分数的解释是相同的：它们依然在平均分之下，它们依然在平均数1.5个标准差下。实际上，你没有改变它，只是改变了用来描述它的数字。

3.5.5 生效原理

z分数的分布为：平均数是0，标准差是1。这是由我们使用的公式决定的。用一组值除以它

们的标准差后，新分布的标准差是1。用分布中的每个分数减去平均数，生成的新值分布在平均数0的附近。

如果我们希望使用的分数有我们自己选择的独特的平均数和标准差，可以对每个z分数进行反转处理，用任何我们偏好的值替代平均数0和标准差1。

3.5.6　理解常模参照计分

我们已经从统计学角度讨论了常模参照计分的内在特点和直观吸引力，但它不是产生有意义分数的唯一方法，也不总是最佳方法。

正如“给原始分数改头换面”[Hack #26]中讨论的一样，当你设计计分系统并建构测试时，实际上有两种原理可供选择。

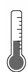
- 常模参照计分

驱动原理：为了更好地理解任务表现（比如参演一部电影或是参加ACT测试），应该对比某个人和其他人的表现水平。

- 标准参照计分

基于一系列标准来评估表现，比如知识库、一套技能、教育性目标和诊断特征。

如果你认为常模参照方法是合理的，那么你就会想用这里介绍的工具来解释自己在这些常见标准测试上的表现。


HACK #28

3.6　正确提问

如果你是一名任课教师、一位面试官，或处于任何想要测量他人理解力的情境下，那么你有多种提问方法。下面是一些测量学工具，能让你以正确的方式提出恰当的问题。

一百多年来，课堂一直是充满问题和答案的地方。除了学校，测试在工作和招聘中也越来越常见。甚至业余时间当我在聚会上遇见他人时，如果不回答我是“友好”还是“冷漠”的关系小测试，我都无法举起一杯Cosmo鸡尾酒。（我是冷漠的，想用它做点什么吗？）

很多教授必须提出好的问题或编写出好的测试：

- ❑ 教师在授课或一对一教学中会对学生提问，以此评估学生的理解程度；
- ❑ 培训师编写问题来评估研讨会的效果；
- ❑ 人事部主任开发标准问题来测量应聘者的技能。

评估他人的学识时，几乎所有人都会面临这样的困境：问哪种类型的问题能真正切中要害。当编写测试或设计问题来测量知识或理解程度时，会遇到两个最常见的问题，而本Hack提供了解

决方案。

- ❑ 如何构建一个好问题？
- ❑ 应该问什么？

3.6.1 构建一个好问题

为了快速且高效地测量知识，很难避免把选择题作为一种问题形式。



多选题（Multiple-choice question）是一种给回答者提供问题或指导语（叫做**题干**），然后让他们选出正确答案或是从一系列答案选项中作出选择的题目。这种类型的题目要求人们选择（select）答案，所以有时候也被称作选项（selection item）。

为了更规范准确地编写好的选择题项，我们使用下面的例子快速入门。

这是一个选择题的例子：

谁写了《了不起的盖茨比》	选　　项
A. 福克纳	干扰项[1]
B. 菲茨杰拉德	正确答案（参考答案）[2]
C. 海明威	干扰项[3]
D. 斯坦贝克	干扰项[4]

如你所见，这个问题的每个选项都有一个名字。正确的答案称作*正确答案*（那怎么能算科学术语呢），错误的答案称作*干扰项*。

对选择题项的特征以及如何编好题项的研究并不是很多，但有一些实证研究。为了编写出好的选择题项，要遵循下面这些通过研究得出的关键项目编写指导原则。

- 包含3~5个选项

题目应该有足够量的答案选项，这样使猜测答案变得困难。但选项不能太多，否则会使干扰项看起来不可信或占用太多答题时间。

- 不要将“以上所有选项”作为选项

有些人会猜测此种选项为正确答案，并将其作为应试策略的一部分。而其他人会避免这种策

注1：Faulkner，1897—1962，美国小说家，曾获1949年诺贝尔文学奖。——译者注

注2：Francis Scott Key Fitzgerald，1896—1940，美国小说家。——译者注

注3：Ernest Hemingway，1899—1961，美国小说家，曾获1954年诺贝尔文学奖。——译者注

注4：John Ernest Steinbeck，1902—1968，美国作家，曾获1962年诺贝尔文学奖。——译者注

略。不管哪种方式，作为一个干扰项，这样操作都是不合理的。而且，评估“以上所有选项”是否为正确答案需要应试者的分析能力，而不同应试者的此种分析能力也是各异的。此外，测量这种特殊的分析能力可能并不是测试的目标。

- 不要将“以上选项都不是”作为选项

这个指导原则的存在原因和上一个指导原则一样。此外，出于某个原因，教师们的确往往把“以上选项都不是”作为最可能是正确答案的选项来设置，有些学生知道这一点。

- 使所有选项可信

如果一个选项看起来和其他选项都不相关，而且明显可以看出它不是正确答案，那么这个选项或许来自测试未覆盖的内容，或许是教师出于幽默原因而将其加入，这样的选项不能作为干扰项。学生不会考虑这个干扰项，所以有4个选项的问题其实只有3个选项可供选择，这样猜中答案就变得更容易了。

- 对选项进行逻辑排序或随机排序

有些教师有这样一种倾向：编写题项的时候让某个答案选项（比如B或C）是正确答案。学生可能会在特定的教师那学到这点。此外，一些用于提高选择题测试成绩的培训课程建议将这一点作为一种应试策略。教师可以通过把选项基于某个规则（比如，从最短到最长、按字母、按时间先后排列）进行排列的方法来控制自己的倾向。



排序问题的另外一个解决方案是：教师在他们的文字处理器上滚动测试的初稿，尝试对选项随机化处理。当然，对于商业标准化测试开发人员来说，计算机随机化也是一个解决方案。

- 使题干长于选项

如果阅读主体在题干，随后紧跟简短的选项，那么答题速度会变得更快。



长题干后跟着短选项，使参加测试的人员处理起来更加容易，一个好的选择题项看起来应该是这样的：

==

======================

======================

======================

======================

- 不要使用否定词

有些学生比其他学生读得更仔细或在文字处理上更准确，但“不是”（not）这个词还是很容易被忽视。即使这个词被强调到每个人都不会忽视它，但教育内容往往不应该作为非事实或错误陈述集来习得，而应该作为积极的措辞真相来存储。

- 让选项和题干语法一致

比如，如果题干中使用的语法很清楚地表明正确答案是女性或是复数，那确保所有的选项都是女性或复数。

- 使用整句表述题干

如果一个题干是完整的以问号结束的问题，或是一个完整的以句号结尾的指导语，那学生能在检验选项之前就开始识别答案。如果题干是以空白或冒号结尾，或者说它只是一个不完整的句子，那么学生需要花费更多的精力来处理此题目。而更多的处理提高了错误的几率。

3.6.2 在正确水平上提问

创建测试时必须要克服的第二个主要问题是：确认所问问题的正确水平。有些问题是简单的，它们只评估某个人的信息再认能力，这种能力代表非常低的知识水平。其他问题更难一点，需要答题者结合现有知识或是将其应用到新问题或情境中。因为不同水平的问题测量不同水平的理解力，如果想从企业获得有用的东西，就必须在正确的水平上提出正确的问题。

有一个聪明的教育研究人员，名叫本杰明·布鲁姆（Benjamin Bloom），他在20世纪50年代提出了一种思考问题的方法，以及正确回答问题所需的理解水平。他的分类体系后来发展成有名的“布鲁姆分类法”（Bloom’s Taxonomy），是一种基于达成某种成就或掌握某种技能所需理解水平的教育目标分类体系。布鲁姆和他的同事给出了学习过程中6种不同的认知阶段。按顺序由低到高排列，分别是：

(1) 知识

词汇、事实和概念的回忆能力；

(2) 理解

理解话题和交流话题的能力；

(3) 应用

使用广义知识解决不熟悉问题的能力；

(4) 分析

将观点分解并理解它们之间关系的能力；

(5) 综合

从已有知识创建一个新模式或观念的能力；

(6) 评估

对新观念的价值作出有根据的判断的能力。

1. 选择正确的认知水平

我们以教师为例，说明如何分析你想问的问题的水平。教师为课堂目标选择合适的认知水平，质量评估的目的是衡量这些课堂目标的达成程度。教师编写的大多数项目，以及那些课本、教材自带的预先编好的测试，都处于*知识*水平。大多数研究人员认为这是不成功的，因为课堂目标的认知水平应该（总是）高于简单记忆信息所需的认知水平。

当新教材被引进时（从学前到高级专业训练的任何阶段），至少要评估是否从中学到了基本的新知识。当教师决定不仅仅测量知识水平时，对项目合适水平的选择取决于学生的发展水平。学生的认知水平，尤其是抽象思考和理解的能力，以及他们使用多个步骤解决问题的能力，决定了课堂目标的最佳水平，因此，也决定了测试项目的最佳水平。研究人员认为，教师应该以他们教课的方式，来测试他们所教授的内容。

所以，任何时候，只要你发现你想评估藏在某人脑袋中的知识，就想想你希望评估的理解力水平。基本的记忆性知识足够吗？如果足够的话，那么*知识*水平就是问题的合适水平。你想知道应聘者是否能够使用他的知识来解决他从未遇到过的问题吗？那就在*应用*水平上进行提问，他不得不证明他是否具有那种能力。

2. 在不同认知水平设计问题

遵循表3-5的指导原则，在布鲁姆分类法的每个水平创建项目或任务。

表3-5：不同认知水平的问题

布鲁姆水平	问题特征	问题或任务示例
知识	只需要死记硬背能力，例如回忆、再认和复述技能	谁写了《了不起的盖茨比》； A. 福克纳；B. 菲茨杰拉德；C.海明威；D. 斯坦贝克
理解	需要释义、归纳和解释等技能	什么是卷尾
应用	需要运算和解决问题等技能，包含*使用*、*计算*和*产生*的词语	如果一个农民原来有40英亩地，又买了16英亩地，那么现在他有多少英亩地
分析	需要列提纲、听、逻辑和观察等技能，包含*确认*和*分解*的词语	画出你邻里的地图并确认每一家
综合	需要组织和设计的技能，包含*对比*和*比较*的词语	基于你对人物角色的理解，描述《献给阿杰尔农的花》（*Flowers for Algernon*）的续集会是怎样
评估	需要批判和形成观点等技能，包含*支持*和*解释*的词语	哪位音乐电影表演家可能是最佳运动员？解释你的答案

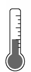
3. 布鲁姆分类法的适用范围

布鲁姆分类法暗含一个等级关系：知识代表认知的最简单水平，评估代表认知的最高和最复杂水平。任何通过编写问题来评估知识的人能够在任意给定水平上编写项目。教师能够确认所选课堂的目标水平，创建和此水平匹配的评估。利用客观计分的项目形式，非常容易达到布鲁姆分类法的低级水平，在更高水平上进行测量会难一些，但也并非不可能。

你不必对布鲁姆定义的6个水平之间的细微差别有太多担心。比如，理解和应用通常被看作同义词，因为应用是指应用所学知识的能力，而这种能力也意味着理解。现如今，大多数测试理论家和任课教师都非常关注知识水平和其他水平之间的差别。大多数教师，除了在全新领域的引入阶段，都更偏爱于教授和测量高于知识水平的目标。

3.6.3 参阅

- 这是我和其他同事合写的学术论文： Frey, B.B., Petersen, S.E., Edwards, L.M., Pedrotti, J.T., and Peyton, V. (2005). “Item-writing rules: Collective wisdom.” *Teaching and Teacher Education*, 21, 357-364（中文书名《教学与教师教育》）。
- 项目编写规则回顾，可参阅：Haladyna, T.M., Downing, S.M., and Rodriguez, M.C. (2002). “A review of multiple-choice item-writing guidelines for classroom assessment.” *Applied Measurement in Education*, 15(3), 309-334。
- 具有影响力的布鲁姆分类法介绍： B.S. (Ed.). (1956). Taxonomy of educational objectives: *The classification of educational goals. Handbook 1. Cognitive domain*. New York: McKay。
- Bloom, B.S., Hastings, J.T., and Madaus, G.F. (1971). *Handbook on formative and summative evaluation of student learning*. New York: McGraw-Hill。
- Phye, G.D.(1997). *Handbook of classroom assessment: Learning, adjustment, and achievement*. San Diego, CA: Academic Press。



HACK #29

3.7 公平测试

任课教师经常创建他们自己的测试来测量学生的学习情况。他们总是担心测试是否太难或太简单，是否测量到了他们想要测量的东西。项目分析工具提供了教师关心问题的解决方案。

课堂评估可能是如今的教室里最常见的活动。教师总是编制测试并给测试评分，学生总是为测试而学习并参加测试，整个过程都是为了支持学生的学习。测试不应该太难（或太简单），并且测试必须测量教师想测量的东西。测试分数和评级是教师和家长、学生、管理人员的沟通方式，所以位于试卷顶部的分数要公平。分数必须准确反映学生的学习情况，并且分数应该是质量评估

的结果。

忧心忡忡的教师一直为改进他们的测试而努力，但是他们总是没有坚实的数据支持，不断在黑暗中摸索。一位聪明且体贴的教师可以通过什么来改进他的测试或提高他评分的效度呢？名为项目分析（item analysis）的一族统计方法能给正在找寻开发公平评估和评分方法的教师指明方向。

3.7.1　项目分析

项目分析是检验个体测试项目课堂表现的方法。一位任课教师也许想检验他编写的测试的部分表现，看他的学生掌握了哪些方面，而哪些方面需要多加复习。一名为护士资格证编制测试的商业测试开发人员也许想要知道他编制的测试中，哪些项目是有效的，哪些项目看起来测量的是其他事物，应该被移除。

在上述两种情况下，测试开发人员会对项目难度和项目效度感兴趣。虽然其中一个例子说的是一位为学生编制测试的中学教师，而另一个例子说的是一个大型盈利公司，但这两类测试的开发人员都对相同类型的数据感兴趣，都能运用相同的项目分析工具。

3.7.2　课堂评估问题的三种类型

如果你是一名担心自己的评估的任课教师，你可能需要回答三种不同类型的问题。幸好，有三种项目分析工具会给你提供三类不同的所需信息。

1. 测试问题是否太难

任何特定测试问题的难度，都可以通过难度指数（difficulty index）公式非常容易地计算出来。你可以计算参加测试的学生中答对题目的人数比例，从而计算某个测试题目的难度指数。比例越大，知道题目所测信息的测试参与者越多。



难度指数这个术语是与我们的直观理解不同，因为它实际上反映的是题目的简单程度，而不是题目的难度。难度指数高的题目是一个简单的题目，而不是一个困难的题目。

多难算难？你得自己决定。有些教师把难度指数为0.50或0.50以下的题目视作太难题目，因为大多数人没有答对。你也许有更高的标准。如果你认为大多数学生应该已经学会了这些内容，而相应题目的难度指数显示班里很大比例的学生答错了，那这个题目可能太难了。

2. 实测是否为想测

测量学家宣称，如果一个测试项目测量的是它想测的东西，那么它就是有效的（valid）[Hack #32]。辨别力指数（discrimination index）是对项目效度的基本测量，此外还要对项目进行信度测量。辨别力指数测量的是项目在整个测试中区分得分高的人和得分低的人的能力。

虽然计算过程有好几步，但计算出来以后，就可以将这个指数看作一个程度指标：反映整个内容领域的知识或技能掌握和项目响应的关系程度。



辨别力指数的得名不是因为它代表测试偏差。辨别力是确认在一个项目上回答正确的人是属于高分组还是低分组的能力。

3. 为什么我的学生错了一道题

除了检测整个测试项目的表现，教师们也对检验选择题的个别干扰项（不正确选项）的表现感兴趣，这种检验是通过选项分析来完成的。通过计算选择每个选项的学生比例，教师能看到学生犯了哪种类型的错误。他们是否理解错了某些概念？他们对资料是否有相同的困惑？

从测量学角度看，为了改进项目的效果，教师们应该确认哪些干扰项是有效的，看起来对那些不知道正确答案的学生有干扰作用；哪些干扰项只是占用一个选项位置，很多学生都不会选择它。

为了消除纯属偶然的、仅靠猜测就答对题目的现象，教师和测试开发人员要设置尽可能多的可信干扰项。对反应选项进行分析，教师能够调整、改进这些也许会在今后的课堂里再次使用的项目。

3.7.3　进行项目分析并解释结果

下面是项目分析的计算方法，我们以带有数据的示例项目进行说明。在此例中，想象有一个课堂，共25名学生参与了表3-6中项目的测试（要记住，即使是标准化测试开发人员对更大规模的、成百上千的测试参与者进行测试，也使用相同的方法）。



表3-6中选项旁的星号代表B是正确答案。

表3-6：项目分析示例

“谁写了《了不起的盖茨比》”一题的答案	选择每个答案的学生数量
A. 福克纳	4
B. 菲茨杰拉德*	16
C. 海明威	5
D. 斯坦贝克	0

为了计算难度指数：

(1) 计算获得正确答案的人数；
(2) 除以参加测试的总人数。

在表3-5的示例中，25人里有16人获得了正确答案：

$$16/25=0.64$$

难度指数范围是0.00~1.0。在我们的例子中，项目难度指数是0.64。这意味着64%的学生知道正确答案。

如果一位教师认为0.64太低了，那么他有一系列措施可供采取。他可以改变他的教学方式来更好地满足项目所代表的教学目标。另一个解释可能是项目太难了，或具有迷惑性，或者无效，在这种情况下，教师可以使用项目辨别力指数或反应选项分析的信息来替换或修改项目。

为了计算辨别力指数：

(1) 按照总分对测试成绩排序，创建两个组：高分组（由排序结果的上半部分构成）和低分组（由排序结果的下半部分构成）；
(2) 对每一组，计算项目的难度指数；
(3) 用高分组难度指数减去低分组难度指数。

假设在我们的例子中，高分组中的13个学生（或测试）里有10人，低分组中的12个学生有6人，答对了本题目。高分组的难度指数是0.77（10/13），低分组的难度指数是0.50（6/12），所以我们能够像这样计算辨别力指数：

$$0.77-0.50=0.27$$

该项目的辨别力指数为0.27。辨别力指数范围为–1.0~1.0。正值越大（越接近1.0），总体测试表现和这个项目表现的相关性越强。

如果辨别力指数是负的，那意味着出于某种原因，测试总分低的学生更可能答对这道题。这是一种奇怪的现象，表明题目效度很糟糕，或者参考答案是错的。教师总是希望测试上的每个项目都是有效的，能反映知识和技能的掌握程度。



计算辨别力指数的公式决定了：如果高分组选择正确答案的学生数多于低分组，那么数字就是正的。所以，至少教师会希望出现正值，因为那将意味着获得正确结果是因为掌握了知识，而不是靠猜测。

我们能利用表3-6提供的信息，来看不同选项的受欢迎程度，如表3-7所示。

表3-7："谁写了《了不起的盖茨比》"项目分析

答　　案	选项受欢迎程度	难度指数
A. 福克纳	4/25	0.16
B. 菲茨杰拉德*	16/25	0.64
C. 海明威	5/25	0.20
D. 斯坦贝克	0/25	0.00

选项分析结果显示，没有答对这道题的学生可能选择了答案A或答案C。没有学生选答案D，所以选项D并没有充当干扰项。在这个项目上，学生不是在4个答案选项中选择，实际上只是在3个选项中选择，因为他们甚至都没考虑选项D。

这样一来，猜对的可能性就更大了，从而降低了项目的效度。教师可能将这个数据视为如下现象的证据：大多数学生在《了不起的盖茨比》和菲茨杰拉德之间建立起了联系，没建立起这种联系的学生无法很好地区分福克纳和海明威。

3.7.4　对项目分析和测试公平性的建议

为了改进测试质量，可利用项目分析确认出太难（或太简单，如果教师有这种担心的话）的项目，但无法区分出掌握内容的学生和没掌握内容的学生，或者说存在不可信的干扰项。

如果作为一名教师，你关心测试的公平性，那么你可以改变教学方式，改变测试方式，或是改变评级方式。

- 改变教学方式

如果有些项目太难，那么你可以调整教学方式。你可以着重教授未学会的内容或者采用不同的教学策略。你也许能明确调整教学方法来纠正学生对内容的困惑和误解。

- 改变测试方式

如果项目有低的或负的辨别值，那可以把它们从当前测试中移除，也可以在将来的测试中将它们从项目库里移除。你同样可以检验项目，尝试发现它的不妥之处，然后改变该项目。当干扰项被确认为无效（没人选择它们）时，教师能够改进项目并创建一个新的干扰项。有效和可信测试的一个目标是降低学生随机猜测出正确答案由此获取分数的几率。可信干扰项的数量越多，测试通常越准确、越有效、越可信。

- 改变评分方式

你可以使用项目分析信息来判断哪些内容是没有教授的，为公平起见，从当前测试中移除该项目，并重新计算分数。对任课教师而言，最简单的做法是：计算出一个测试中的不良项目数，

并将这个数字加到每个学生的分数上。此方法与把这些项目当做不存在而重新计分的方法在技术上是不同。但是这样的话，学生如果答对了某个难度较高的项目，他们依然能够获得该项目的分数，对大多数教师来说，这种方法看起来更加公平。

这些教师对测试质量的关心和科学家提出的研究问题没有太大差异。就像科学家一样，教师可以在他们的课堂上收集、分析数据并解释结果。他们能够基于自身的认知体系，决定对结果采取什么样的措施。



3.8 什么都不做也能提高测试分数

如果你对刚刚参加过的一场利害攸关的测试分数不满意，也许你应该再次参加这个测试。你觉得呢？

我们已经讨论了如何运用信度[Hack #6]的概念来精确测量事物。信度是指测试评估结果的一致性。换句话说，可信的测试产生稳定的分数，不可信的测试无法产生稳定的分数。由于测试不是完全可信的，所以其产生的分数至少有一部分是有随机性的，这些分数按照统计学家预测的方式上下波动。因为当你再次参加测试时，你的分数往往在测试的平均分上下波动，所以这种效应叫做*均值回归效应*（regression toward the mean）。

当你参加一些利益攸关的测试时，比如SAT、ACT、GRE、LSAT或MCAT，你总是有重新参加测试来尽力提高分数的机会。关于是否值得花时间、精力和金钱去尝试提高你的分数，取决于对测试信度的理解以及仅通过简单的均值回效应来提高分数的可能性。

3.8.1 均值回归

首先，让我们制造出一个均值回归，这样你就会相信，只因为正态曲线[Hack #23]特征，分数就能在预测方向上改变。眼见为实，我希望在你眼前呈现这个无形的神奇现象。

把表3-8中的判断题发给你关系最好的100个朋友。嗯，好吧，也许包括你在内有10个人也行。1000个会更好，只要数量足以让我向你证明回归的发生即可。我们准备这项测验时要记住，如果有100个或是1000个人参加这个非常难（或简单）的测试，那结果会更令人信服。

对于这个测试，你不需要看实际的问题。这个测试的测量内容和结构[Hack #32]不需要任何改变，分数就会改变。所以，在这个小测试上，你所能做的就是猜。因为它们是判断题，所以每道题你都有50%的几率答对。你那10人测试组（或是100人，如果你真的很在意这点……你能够至少找到30人吗？……还有谁愿意参加）的平均分应该是5分（满分为10分）。

表3-8：高等量子物理小测试

问题	圈出你的答案
1.	对或错
2.	对或错
3.	对或错
4.	对或错
5.	对或错
6.	对或错
7.	对或错
8.	对或错
9.	对或错
10.	对或错

让你所能联系上的所有人都来参与这个高等量子物理测试。当你和其他人参加这个测试时，即使标准答案近在咫尺（在表3-9里），也不要作弊去看标准答案！

表3-9：高等量子物理测试标准答案

1. 对	2. 对	3. 错	4. 错	5. 对
6. 错	7. 错	8. 对	9. 对	10. 错

把这些完成的测试（确保他们都填写了名字）收集上来，然后对照表3-9中的答案进行计分。

现在，选出你小组的得分最高者（这可能代表某些和你一样的人，他们在标准测试，比如SAT中，得分高于平均分），得分最低者（这也许表示某些和你不一样的人，他们的得分比平均分低）。对这两个人再次进行测试（他们没有查看正确答案），然后再次计分。

下面均值回归起作用了。不需要了解你或你的朋友，也不需要知道他们的答案是什么，有两件事情我相当肯定：

- 第一次得分最低的那个人，第二次的得分会比他第一次的高；
- 第一次得分最高的那个人，第二次的得分会比他第一次的低。

如果真是如此，那么啊哈！我早就告诉你了嘛！如果不是这样，我跟你说的只是“相当肯定”而已。如果有更大的样本，结果如此的可能性也更大。

3.8.2　生效原理

我们对这两个分数的预期是：所有低于5分（或是你测试的平均数）的测试分数会向上移动，趋向平均数；所有高于5分的测试分数会向下移动，趋向平均数。你的两个分数可能会出现这种

情况，也可能没有出现这种情况，但它是最可能的结果。

记住，这是一项知识对分数没有影响的测试。两次分数都完全是由几率导致的。但是，即使在知识会影响分数的真实测试中，这种效应也会出现。那是因为没有一个真实测试是完全可信的，几率在每个测试上多少会起点作用。这个例证只是将测试置于几率百分百地影响测试人员分数的情境下，由此夸大了这种效应。

那么，为什么在第二次测试时，分数可能发生改变并向平均数靠拢呢？从长远来看，有100个或1000个测试分数集合，我们会期望某种像正态分布的结果。就像扔硬币一样（结果可以是正面或反面，两种情况下的几率都是50%），在判断题测试上（或任意测试），概率都是和特定的结果联系在一起的。表3-10展示了高等量子物理测试中，可能的分数以及测试人员得到那个分数的概率。

表3-10：可能的测试分数分布

分　数	概　率
0	0.001
1	0.010
2	0.044
3	0.117
4	0.205
5	0.246
6	0.205
7	0.117
8	0.044
9	0.010
10	0.001

为什么很极端的分数在重复测试后变得不那么极端了？看看得到两个极端分数（比如第一次是2分，然后第二次也是2分）的概率，对比第一次分数是2分（概率是0.044），然后第二次是4分（概率是0.205）的概率。一个人第一次得2分，第二次得4分的概率几乎是两次都得2分概率的5倍。几乎有95%的把握说他会获得高于2分的分数（1−0.044−0.010−0.001=0.945）。



"均值回归"一词得名于著名的弗朗西斯·高尔顿（查尔斯·达尔文的堂弟），他研究父母和成年子女的身高问题。他发现，成年子女的平均身高更接近于所有成年子女的平均身高，而不是他们父母的平均身高。虽然高尔顿把这个观察结果称作"平庸回归"（由此高尔顿不再仅仅作为外交家而知名），但我们会友善一点。这和遗传没有任何关系，但和统计有密切联系。

这个测试的分数完全受几率影响，有65.6%的几率能得到平均数或非常接近平均数的分数（4

分、5分和6分的组合概率）。对于大多数测试来说，它们有着更多的题目数，形成正态分布，这样你有68%的几率获得平均数或接近平均数的分数[Hack #23]。

3.8.3 预测获得更高分数的可能性

有趣倒是有趣，但它如何帮助你判断是否值得再次参加测试呢？这就回到了我们最初的两难问题上。再次参加这些重要测试（比如大学录取测试），会花费更多的钱、时间并带来更大的压力，也许还需要准备，所以我们需要战略性地决定什么时候再次重试。



当然，你可以通过提高测试所需的知识水平来真正提高自己的考试成绩。如果你通过学习、参加模拟考试或预备课程等来准备测试，你可能会获得更高的分数。但是，如果你获得很低的分数，那么即使在两次测试间隔期你什么都不做也有可能提高分数，就因为均值回归。你能在两次测试间隔期很轻松，而分数依然可能提高。真是幸运儿！

只通过再次参加测试，你就能获得更高的分数，这种可能性取决于两件事：你第一次的测试分数和测试的信度。

- 你的分数

因为分数可能（只因为几率）向平均数移动，给你第二次机会，你能做得更好的几率取决于你第一次的分数究竟低于还是高于平均数。把平均数想象成你听见的巨大吮吸声，它将所有的分数沿着分布拉向它。平均数以下的分数比平均数以上的分数更有可能上升。

- 测试的信度

测量统计学家用一个数字表示信度，代表并非由几率导致的分数变异比例。那么，信度越高，几率在决定你分数时起的作用越低。可信分数是稳定的分数，平均数的超级吮吸力不如一个可信分数。

统计学家已经开发出了一个公式，你能运用这个公式计算分数的变化空间。如果有足够的成长空间，你可能考虑第二次尝试。这里用到的一个非常有用的工具是测量标准误差。下面是测量标准误差[Hack #6]的公式。

$$\text{标准误差}=\text{标准差}\sqrt{1-\text{信度}}$$

大多数标准测试在每次执行期间，都会发布由测试产生的成千上万个分数的信度水平和期望标准差。通过将这些测试的值代入测量标准误差的方程式里，会对从测试到重测的分数变异有一个大致概念，这种变异可能在被测人员没有任何真正改变的情况下发生。

但是，即使是标准误差，对极端分数来说，也可能出现误导。非常低的分数和非常高的分数，仅由几率导致的移动距离可能比标准误差建议的距离要大。你离正态越远，抗拒正态分布的万有引力就越难。极端分数无法抗拒那种引力，除非它们是完全可信的。

总之，下面是关于如何决定是否该重新参加测试的合理建议。

- 如果你获得了相对很高的分数，但没有高到你期望的水平，那么可能不值得再参加一次测试。
- 如果你获得的分数很低（远低于平均），几乎可以肯定你第二次的分数会更高。再试一次吧。第二次你应该也更努力了一些。

——尼尔・萨尔金德



3.9　建立信度

对使用、编制和参加利益攸关的测试的人而言，建立测试分数的准确性是很有利的。幸好，教育和心理测量领域提供了几种方法可以验证测试分数的一致性、准确性，并表明其可信程度。

任何使用测试来进行重大决策的人，都需要确定产生的分数是准确的，并且分数没受到太多随机作用的影响，比如那天早晨的应聘者是否吃了早饭，或学生在测试期间是否过度紧张。测试开发人员需要建立信度来说服他们的客户相信可以依赖产生的结果。

也许，最重要的是，当你参加一项关乎能否被录取，或决定是否会晋升为首席餐饮大厨的测试时，你需要知道分数反映了你的典型水平。本Hack展现了信度测量的几种方法。

3.9.1　信度的重要性

首先，讲解一些关于测试信度的基础知识，以及你为什么要找出你所参加的重要测试的信度证据。人们期望测试和测量工具有一致性，不管是内部的（用相似方法测量相同的构造行为）还是外部的（如果横跨不同时间反复执行，那么得到相似的结果）。这些都是信度的问题。

信度通过统计方法来测量，可以通过计算一个特定的数字来代表一个测试的一致性水平。大多数信度指标基于如下相关[Hack #11]：对测试项目做出的反应之间的相关，或一个测试的两个分数集之间的相关，或是一个测试两次计分的相关。

有四种常见的信度类型用来确立一个测试产生的分数是否不包含太多随机变异：

- 内部信度

每个参试者的表现在同一个测试中的不同项目间是否一致？

- 重测信度

执行同一测试两次，每个参试者的表现是否一致？

- 内部评分者信度

如果两个不同的人给测试评分，参试者的表现是否一致？

- 平行信度

采取不同形式执行同一个测试，参试者的表现是否一致？

3.9.2　计算信度

如果你已经编制了一个你想使用的测试——不管你是一名教师、一位人事主管还是一位临床医学家，你都想证实你的测试是可信的。你用来计算准确性水平的方法取决于你感兴趣的信度类型。

1. 内部信度

最常见的信度测量是内部一致性测量，也称作 α 系数（或克隆巴赫系数）。系数α是一个几乎总是介于0.00~1.00的数字。值越大，测试项目的内部一致性越高。

如果你参加一个测试并把测试分成两半，比如奇数项为一半，偶数项为另一半，你能计算出这两半的相关性。计算“分半相关”（split-half correlations）的公式就是计算相关系数的公式[Hack #11]，并且计算分半相关是一种常用的估计信度的方法，虽然分半信度现在被认为有点过时。

从数学上讲，计算系数 α 的公式产生了一个测试所有分半可能的平均相关，并且已经替代了分半相关，成为了估计内部信度的首选。因为这个方程的计算比较复杂，通常用电脑来计算这个值。

$$\alpha = \frac{n}{n-1}\left(\frac{SD^2 - \sum SD_i^{\ 2}}{SD^2}\right)$$

n代表测试的项目数，SD代表测试的标准差，S表示加总，SD_i 表示每个项目的标准差。

2. 重测信度

内部一致性被认为是代表测试信度的合适证据，但在一些情况下，有必要证明过一段时间后问卷的一致性。

如果被测量的事物随着时间推移不会改变，或者它会缓慢改变，那么，如果在两个不同时间执行相同的测试，相同群体的反应应该非常一致。这样两个分数集合之间的相关会反映测试随着时间推移的一致性。

3. 内部评分者信度

当不止一人观察测试评分时，我们同样能够计算信度。采用不同评分者的评分时可以证明不同评分者的一致性。甚至只有一名评分者（如一位任课教师）时，如果评分是完全主观的，因为大多数题目是问答题和绩效评估，那么这种类型的信度也有很大的理论意义。

为了在这些情况下，证明个体的分数代表典型表现，必须证明即使使用不同的评判员、计分员或评定者，结果也是没有差异的。内部评分者信度水平的确定通常是建立一系列评分者的分数相关性或计算他们意见一致程度的百分比。

4. 平行信度

最后，我们能通过论证下面这个问题来证明信度：一个人参加何种测试的形式不重要，他在这些测试上都会获得相同的分数。只有测试是从大项目池中构建时，才有必要证明平行信度。

比如说，很多标准化大学的入学测试，例如SAT和ACT，不同的参试人员参加不同版本的测试，这些测试是由覆盖相同主题的不同问题构成的。这样的话，即使你周六早晨在缅因州参加了测试，也无法给你在加利福尼亚的堂兄打电话告诉他具体的考题，以便他为下周的考试作准备，因为你的堂兄可能在他考试时遇到一组不同的问题。

当公司编制不同形式的相同测试时，他们必须证明那些不同形式的测试难度相同，还有其他相似的统计属性。最重要的是，他们必须证明，你的缅因州版本的测试分数会和加利福尼亚版本的测试分数相同。

3.9.3　解释信度证据

有多种方法可供建立测试信度，不同目的的测试应该有不同的信度证据。你能根据信度系数的大小来决定你刚刚编制的测试是否需要改进。如果你只是参加测试或只利用测试提供的信息，那你能用信度的值来判断是否应该相信测试的结果。

- 内部信度

只设计用来做重要决策的测试，应该有非常高的内部信度，这样一个人在这个测试中获得的分数应该会非常准确。虽然只是一个经验法则，但人们通常认为0.70或更高的 α 系数是声明一项测试具备内部信度所必需的。对于你来编制或参加的测试，还是你自己决定多大的信度是可接受的吧。

- 重测信度

像很多社会科学研究设计一样，一个用来测量随时间推移发生的变化的测试，应该展示良好的重测信度。良好的重测信度意味着多次测试之间分数的改变不是由随机波动导致的。稳定相关

系数的合适大小取决于随着时间的推移，结构的理论稳定程度。那么，取决于它的特征，随着时间的推移，测试产生分数的相关性介于0.60~1.00。

- 内部评分者信度

内部评分者信度只有在计分受主观因素影响的情况下，比如写论文测试，才会令人关注。客观的、计算机计分的选择题测试应该产生完美的内部评分者信度，所以通常对客观测试来说，不会产生那种类型的证据。如果内部评分者相关被用来估计内部评分者信度，那么根据经验法则，0.80是最小可接受的内部评分者信度水平。

有时，内部评分者信度通过报告两位评分者意见一致性的百分比来估计。用一致百分比来估计时，通常认为比例达到85%就足够了。

- 平行信度

只有存在不同形式的测试才能被描述为具有平行信度。你的大学教授可能不需要建立平行信度，因为期末测试只有一个版本，但是大规模的测试公司可能需要建立平行信度。

平行信度应该非常高，这样人们能将测试的任何形式视作具有同等意义。通常来说，一项测试两种形式之间的相关性应该高于0.90。测试公司采取这样的研究方式：一群人都按照两种形式参加这项测试，以此来计算平行信度系数。

在你参加一项利益攸关的、关乎未来发展的测试之前，确保测试有可接受的信度水平。你希望看到的信度类型证据取决于测试的目的。

3.9.4　改进测试信度

要确保测试有一个高α系数或其他任何信度系数，最简单的方法是增加测试的长度。围绕相同概念进行提问的项目越多，作答者澄清他们态度或展现他们知识的机会就越多，那项测试上总分的信度就越高。这在理论上讲得通，也同样从数学上提高了信度，我们可以从计算信度所用的公式看出来。

回顾前面的α系数计算公式。随着测试长度增加，总测试分数的变异比项目间的总变异增长得更快。在公式里，这意味着随着测试变长，括号里的值变大。$n/n-1$部分同样随着项目数量增加而提高。所以，更长的测试往往产生更高的信度估计。

3.9.5　生效原理

相关性使两个分数集匹配起来，每对分数描述一个个体。如果多数人表现一致——两个分数都高或都低，或者和其他人对比都是平均水平，或者一个测试的高分与另一个测试的低分匹配一致，那么相关性会接近1.00或−1.00。

分数之间的不一致关系，产生一个接近于0的相关。分数的一致性，或是测试和其自身的相关，在经典测试理论[Hack #6]建立的标准下，可表明分数是可信的。经典测试理论认为，除了其他方面外，随机误差是单人多次参加相同测试而分数发生变化的唯一原因。



3.10 建立效度

一项测试最重要的特征是，它对预期目的有用。如果要证明测试分数代表了预期设定的意思，那么建立效度是非常重要的。如果你可以提供某种类型的证据，那么能够让你自己和其他人相信你的测试是有效的。

一个良好的测试测量它打算测量的事物。比如一项意图找出高中生系汽车安全带频率的调查，很明显，这项调查应该包含关于安全带使用的问题。一个没有这些项目的调查，会因为没有效度而受到合理批评。调查、测试和实验都需要可接受的效度。如果你正在设计一项心理学或教育测试，或只是想确保你的测试是有用的，那么你应该关心效度的建立问题。

对一个测试而言，效度不是可有可无的东西。效度是由测试开发人员、那些关心测试结果的人，以及任何与测试及测试结果利益相关的人共同决定的。

想想一个由数学问题构成的拼写测试。很明显，数学问题构成的测试不是一个有效的拼写测试。虽然它不是一个有效的拼写测试，但它可能是一个有效的数学测试。测试的效度或调查的效度不在于工具本身，而在于对结果的解释。

一项测试可能对一个目的有效，但对另一个目的无效。用一个学生的拼写测试分数来解释他的数学能力是不合适的。这个分数也许作为对语言能力的测量是有效的，但对数字流体能力[5]（fluidity）无效。分数本身既不是有效的也不是无效的，与分数关联的意义才是有效或无效的。

为了说明如何解决建立效度的问题，想象你设计了一种测量拼写能力的新方法。你想要把测试卖给全国的学校，但首先你必须拿出显而易见的证据，证明你的测试测量的是拼写能力，而不是其他内容，比如词汇、焦虑性、阅读能力或是（其他可能影响分数的因素）性别或种族。

3.10.1 效论的制胜策略

效度看起来像一个永远无法获胜的辩论，因为作为一个不可见的质量指标，它永远无法完全建立起来。但作为一名测试开发人员，你希望使参试人员以及任何会使用测试结果的人相信，你本质上测量的就是你想要测量的事物。幸好，有很多可行方法能够给测试提供效度证据。

注5：在心理学的智力领域，美国心理学家卡特尔把智力分成流体能力和晶体能力，流体能力是人的一种潜在能力，主要和神经生理的结构和功能有关，很少受社会教育影响，它与个体通过遗传获得的学习和解决问题的能力有联系。晶体智力则主要是后天获得的，受文化背景影响很大，与知识经验的积累有关。——译者注

有趣的是，最普遍接受的效度类型在理论上具有最弱的论据。这种论据是表面效度的一种，它是这样的：测试是有效的，因为它看起来（表面上）像测量了它想要测量的事物。那些提出或接受表面效度论据的人认为，在这个测试中发现了他们期望的项目类型。比如，之前提到的安全带使用调查，如果其中有项目问到安全带使用，那么它就会被视为具有效度。

表面效度论据很弱，因为它只依赖于人们的判断，却令人无法抗拒。在说服某人完全相信并接受一个评估时，常识是一个很强的论据，甚至可能是最强的。虽然表面效度看起来没有其他类型的效度那么具有科学性（实际上，它是不太科学的），但如果缺少表面效度，那些编制者和使用者几乎不会接受这种测试工具。作为一名测试开发者或用户，如果你不能提供本Hack后面讨论的效度类型，那么你应该提供一个至少具有表面效度的测试。



对于你的拼写测试，如果参试者被问到拼写问题，就说明你已经建立了表面效度。

有四种更科学的效度证据，被那些经常运用评估的人普遍接受。它们都属于效度的论据范围。

- 基于内容的论据

测试中的项目公正地代表了能在这个测试上出现的项目吗？如果一个测试想要覆盖一些明确界定领域的知识，那么问题是从这个领域公正取样的吗？

- 基于标准的论据

测试的分数能用于估计其他类似测试的表现吗？

- 基于结构的论据

测试的分数代表了你希望测量的特质吗？

- 基于结果的论据

参加测试的人受益于经验吗？测试时偏向于某个群体吗？参加测试是否导致太多的压力，以至于不管分数如何，都是不值得的？

3.10.2 基于内容的论据

假设你决定测量一个概念，而那个概念有很多方面，并且在一个测试上能问很多不同的问题。你需要证明为测试选择的项目代表了所有的可能项目，这种证明就是对效度基于内容的论据。

这听起来像一个令人畏惧的需求。通常，人们认为这类证据在测量成就时更加重要。在成就领域（如医药、法律、英语、数学），有非常多且明确清楚的领域和内容可供某项有效测试取样。同样，一名任课教师可能已经定义了一项测试应该测量的一系列目标或内容范围。但是，当测试行为、知识或态度这些领域时，很难像这样准确定义一个学科的各方面。因此，作出这样一个合

理论据是困难的：你已经选择出了一些问题，它们能代表某个想象的所有可能问题的问题池。

那么，在测试构建中，对效度的内容证据而言，什么才是必须的？看起来，至少需要某种问题选择或构建的组织方法。比如，当测量自尊时，问题可能涵盖参试者在不同环境中的自我感觉如何（如工作场所、家里或学校），同时还有不同任务表现（如体育、学术或工作职责），或对自己不同方面的感觉如何（如外表、智力或社交技能）。



对于一名测量过去几周学生学习程度的任课教师来说，制定一张**规范表**（包含组织好的主题列表并表明重要性）是个好方法。

测试开发人员有权决定如何组织一个概念或如何将这个概念分解。测试人员可能从研究或其他测试中获得灵感，也可能只是遵循了一些通用模式。关键是要说服自己，这样你才能说服他人，让他们相信你的选择覆盖了正测量事物的重要方面。

对于你的拼写测试，如果能证明让学生拼写的单词代表了学生应该掌握的更大的单词池，那你就是在提供基于内容的效度证据。

3.10.3　基于标准的论据

效度的标准证据说明，一个测试上的回答能预测某个其他情境下的表现。“表现”可以是工作上的成功，测试分数、他人的评价，等等。

如果测试上的回答和标准表现相关，且这个标准能马上测量，那么这个效度证据叫做*同时效度*（concurrent validity）。如果对测试的响应和未来某天才能被测量的标准表现相关，那么这个效度证据叫做*预测效度*（predictive validity）。

显而易见，你选择用来支持标准效度的测量应该具有相关性，测量的概念应该与标准具有或多或少的理论相关性。当测试的明确目的是估计或预测在某个其他测量上的表现时，这种形式的效度证据是最具说服力且最重要的。

当测试不需用来预测未来或估计在某个其他测量上的表现时，基于标准的论据就不那么具有说服力了，也许是不相关的。比如，这种证据可能对你的拼写测试没有用。另一方面，你也许可以证明在你的测试中得到高分的人，在全国拼写比赛中也表现良好。

3.10.4　基于结构的论据

效度证据的第三种类别是结构证据。*结构*（重音在第一个音节，con-struct）是一个测试设计要测量的理论概念或特质。我们知道永远无法直接测量智力或自尊等的结构。心理测量的方法是间接的。我们通过问一系列问题，希望作答者使用我们正测量的他思维的一部分，或参考包含过

去行为或知识信息的记忆的一部分，或者，至少指引作答者检验他在某个特定话题上的态度和情感。

我们进一步希望参试人员在测试项目上准确且诚实作答。实际上，测试结果总是被当做结构的直接测量，但我们不应该忘记它们只是有根据的推测。整个过程的成功依赖于另外一系列假设：我们已经正确定义好了我们试图测量的事物结构，并且我们的测试也反映了那个定义。

那么，结构证据总是包含这两方面：对所定义结构本身的辩护和对使用工具反映了定义的声明。展示结构效度的论据包含这样一个论证：实际的反应和理论预期的反应一致。结构效度在每使用一个调查或测试时不断累积，像所有的效度论据一样，它永远无法完全令人信服。在某种意义上，结构效度论据包含了内容和标准效度论据，因为所有效度论据都试图建立概念和测量之间的联系。

对于你的拼写测试，可能存在对拼写能力本质的研究，将其作为认知活动、人格特质或某种其他明确定义的实体。如果你能通过拼写能力定义你的意思，证明你的测试分数和定义所期望的一致，那你就拥有了基于结构的效度证据。理论认为阅读能力好的人拼写能力也好吗？展示那种相关，也许用到相关系数[Hack #11]，这样你就已经呈现了可能说服别人的效度证据。

3.10.5 基于结果的论据

在10年或20年之前，对建立效度感兴趣的测量人员只关心如何证明测试分数反映结构。随着人们开始关注一些测试可能会不公平地使整组人处于不利地位，加上担忧测试的普遍使用会带来社会问题，政策制定者和测量哲学家们现在开始审视参试者因为参加测试而导致的后果。

我们如此习惯测试并基于那些分数进行利益攸关的决策，现在我们应该偶尔退一步，问问自己，如果依赖测试做决策，社会是否会更进步。从代表测试结构的分数到满足预期目的测试，效度的含义在不断扩大。想必测试是在这里给世界提供帮助的，而不是伤害它，基于结果的效度证据是用来证明测试的社会价值的。



就像古老笑话中政府人员一样，测试是“在这给我们提供帮助的”。

对于你的拼写测试，你想要消除的核心负面影响是测试偏差。如果你的拼写能力理论预期性别、种族或社会经济地位之间没有差异性，那么拼写分数在这些组间应该相同。你也许可以用t检验[Hack #17]，来提供组间分数相似的证据，这样就很好地证明测试的公平性和有效性。

3.10.6 从效度菜单选项里选择

这里描述的不同效度证据类别都代表一个策略性的菜单选项。如果你想要证明效度，可以从这些不同的效度证据类型中选择。

明显地，不是所有的测试都需要提供所有类型的效度证据。一项由教师为25个学生编制的小型测试，可能只需要一些基于内容的效度证据来说服教师相信测试结果。基于标准的效度证据不是必须的，因为估计在另一个测试上的表现不是这类测试的预期目的。

另一方面，重要的测试，比如大学入学测试（像ACT、SAT和GRE）和智力测试，这些用来确认学生特殊教育基金资格的测试，应该得到四个效度的证据支持。对于你的拼写测试，你可以自己决定提供哪种类型的证据、哪种类型的论据是最有说服力的。



3.11　预测生命周期

我们中的很多人凭直觉相信，已经存在很长时间的事物可能会存在更长的时间，存在不久的事物，继续存在的时间也不长。这种图式的形式化叫做戈特原理（Gott's Principle），数学上也容易证明。

到目前为止，物理学家理查德·戈特三世（J. Richard Gott III）已经成功地预测出柏林墙的倒塌，计算出44大道百老汇的持续时间。具有争议的是，他预测人类可能存在的时间介于5100年至780万年之间，但不会再长。他认为这是创建自给自足型太空殖民地的一个很好理由：如果人类把卵放到其他巢里，也许可以因此使其避免行星撞击或星球家园的核战争，从而延长我们种族的生命周期。

戈特认为他的简单计算在某些参数范围内，能够被运用到几乎所有事物上。为了使用这些计算来预测某事物的存在时间，你需要知道它已经存在了多长时间。

3.11.1　行动起来

戈特的计算基于他所谓的哥白尼原理（在这个特定应用下，有些人将其称作戈特原理）。这个原理假定，当你选择某个时刻来计算某个现象的生命周期时，那个时刻可能非常普通，不是特别的或是享有特权的，正如哥白尼告诉我们地球在宇宙中并不占据特权位置一样。

在普通的、无特权的时刻选择对象，这一点很重要。选择你认为处于生命周期开始或结束阶段的对象，比如住在新生儿病房或疗养所的人，会让你的测试有偏差，产生糟糕的结果。进一步说，戈特原理在有确切数据存在的情况下，不是那么有用。由于已经有了大量关于人类生命周期的精确数据，所以戈特原理在这个方面也不是那么有用。

假设我们已经选好了一个时刻，现在来检验它。在其他所有条件都一样的情况下，这个时刻处于这个现象生命周期中间的50%，这种情况发生的几率为50%，有60%的几率处于中间的60%，有95%的几率处于中间的95%，以此类推。因此，只有25%的几率你选择的时刻在前1/4的生命周期里，20%的几率在前1/5里，2.5%的几率在生命周期的后2.5%里，以此类推。

表3-11给出了50%、60%和95%的置信区间的方程式。t_{past}表示对象已经存在的时间，t_{future}代表它预计还能存在的时间。

表3-11：戈特原理的置信水平

置信水平	最小t_{future}	最大t_{future}
50%	$t_{past}/3$	$3t_{past}$
60%	$t_{past}/4$	$4t_{past}$
95%	$t_{past}/39$	$39t_{past}$

让我们看一个简单的例子。请快速回答：从现在开始算起，你认为谁的作品更有可能被大家再听上50年？约翰·塞巴斯蒂安·巴赫（Johann Sebastian Bach）还是布兰妮·斯皮尔斯（Britney Spers）？巴赫的第一部作品大约出现在1705年。从我现在写书的时间看，那是300年前了。布兰妮的第一张专辑在1999年发布，大约是6.5年前或79个月前。

查询表3-10，对60%的置信区间，我们看到t_{future}最小值是$t_{past}/4$，最大值是$4t_{past}$。

因为布兰妮音乐的t_{past}是79个月，所以有60%的几率，继续听布兰妮音乐的时间介于79/4个月至79×4个月之间。换句话说，我们有60%信心说，从现在，布兰妮会有介于19.75个月（1.6年）至316个月（26.3年）的文化影响力。



>60%对快速估计来说是一个良好的置信水平，不仅因为它是一个比平均要好的几率，还由于1/4和4容易计算。

出于同样的原因，我们能在60%的置信水平，预期从现在起人们听巴赫音乐的时间介于300/4和至300×4之间，或者说75年至1200年之间。因此，我们能预测，布兰妮的音乐很可能会和她的粉丝一同消亡，而巴赫的音乐可能会一直被听到第四个千禧年。

3.11.2 生效原理

假设我们正在研究我们称作目标的某个对象的生命周期。正如看到的那样，我们有60%的几率处于这个对象生命周期中间的60%（如图3-4所示）。

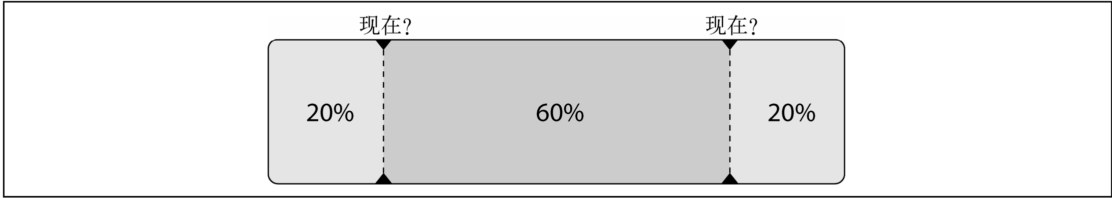


图3-4：生命周期中间的60%

如果我们处于中间60%的最末端，那我们就是在图3-4中标记“现在”的第二个点位置处。在这个点上，目标生命周期只剩下20%（如图3-5所示），意味着t_{future}等于1/4的t_{past}（80%）。这是我们在60%置信水平预期的最小剩余生命周期。

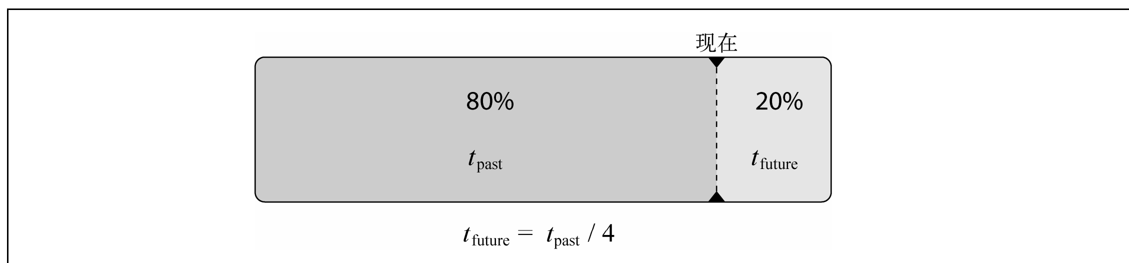


图3-5：最小剩余生命周期（60%置信水平）

相似地，如果我们处于图3-4里中间60%的开始之处（标记“现在”的第一个点），那么未来还有80%的目标生存期，如图3-6所示。因此，t_{future}（80%）等于$4 \times t_{past}$（20%）。这是我们处于目前置信水平预期的最小大生命周期。

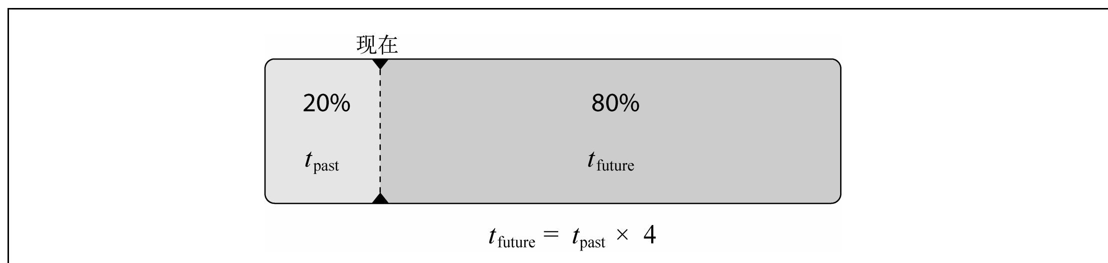


图3-6：最大剩余生命周期（60%置信水平）

因为位于这两点之间的几率有60%，所以我们能以60%的信心算出目标未来的持续时间介于$t_{past}/4 \sim 4 \times t_{past}$。

3.11.3　现实应用

假设你想要投资一家公司，并且想估计这家公司会存活多久以判断这是否是一个好的投资。你能使用戈特原理。让我们以这本书的出版商O’Reilly Media为例，虽然它没有上市。



我当然不是随机选择O’Reilly Media的，关于公司能持续多久有丰富的历史信息，但是让我们至少尝试把戈特原理作为对O’Reilly寿命的一个简易估计。毕竟，对于百老汇演出的寿命可能有很好的数据，但是戈特并不畏惧分析它们。所以我不会说现在O’Reilly已经出版了*Mind Performance Hacks*，它的不朽是肯定的。

由维基百科得知，O’Reilly作为一家从事技术写作的咨询公司，始于1978年。我写作此书的时间为2005年7月，所以O’Reilly作为一家公司已经存在了大约27年的时间。我们预计O’Reilly还将存活多久？

下面是O’Reilly可能的生命周期，置信水平为50%：

- 最小

27/3=9年（到2014年7月）

- 最大

27×3=81年（到2086年7月）

下面是置信水平为60%时的生命周期预期：

- 最小

27/4=6年零9个月（到2012年4月）

- 最大

27×4=108年（到2113年7月）

最后，置信水平为95%时的生命周期预测：

- 最小

27/39=0.69年=大约8个月零1周（到2006年3月中旬）

- 最大

27×39=1053年（到3058年7月）

在互联网经济时代，这些数字看起来相当不错。例如，苹果公司好不到哪儿去，微软成立于1975年，所以我们同样可以这么说微软。一个真正的投资者会考虑很多其他因素，比如年收益和股票价格，但初看之下，好像O’Reilly至少会在下个十年比一个假设的投资者活得更长。

3.11.4 参阅

- Ferris, Timothy. “How to Predict Everything.” *The New Yorker*, July 12, 1999.（中文书名《纽约客》，1999年7月12日出版。）
- Gott, J. Richard III. “Implications of the Copernican Principle for Our Future Prospects.” *Nature*, 363, May 27, 1993. （《自然》杂志，1993年5月27日版。）
- Gott, J. Richard III. “A Grim Reckoning.” http://pthbb.org/manual/services/grim.

3



3.12　作出明智的用药决定

医学检测提供的诊断甄别信息总是容易被病人误解，有时候，医生甚至也会误解。理解“敏感性”和“特异性”的概率特征能提供更加准确和更安心（有时候）的概览。

作为一名医学信息的消费者，你必须对行动、治疗和再找名大夫寻求第二意见等做决定。你可能会依赖于药物信息——新闻故事、你医生的建议、检测结果，来做这些决定。但是，你从医生那获得的大多数药物信息，都有一个已知的误差量。这对指出你患某种疾病的概率的诊断检测结果来说，尤为正确。

本Hack讲解使用医学检测的特征信息来获得更准确的事实，以期能对治疗作出更好的决策。

3.12.1　统计和药物甄别

为了明智地使用医学检测信息，我们必须掌握一些概念的准确意义。用准确性的语言表述医学检测的四种可能结果，如表3-12所示。

表3-12：医学检测的可能结果

	患者真患病（A）	患者实际没患病（B）
检测结果显示患者患病	真阳性（分数是正确的）	假阳性（分数是错误的）
检测结果显示患者没患病	假阴性（分数是错误的）	真阳性（分数是正确的）

医学筛查检测的信度[Hack #6]被概括为敏感性和特异性的这两个比例。本质上说，依赖这些检测结果做决策的人，关心有关准确性的三个问题。

如果一个人患有疾病，这个人得到阳性试验结果的可能性是多少？这种可能性称作敏感性。在A列中的那些人，得到阳性试验结果的概率是多少？

如果这个人没患病，那这个人的检测结果为阴性的可能性是多少？这种可能性称作特异性。在B列中的那些人，检测结果为阴性的比例有多大？

如果一个人的检测结果为阳性，那这个人患病的可能性是多少？从病人角度看，这是一个终极问题，这个问题可以被认为是这类检测关心的基本效度。医生，我能够相信这类检测结果吗？还是说检测结果有错误？



注意在表3-12中，A列和B列是不同的人。患病的人是在A列，未患病的人在B列。如果你在A列，你在检测中不能得到假阳性的结果，因为阳性结果是正确的。如果你在B列，你在检测中不能得到假阴性的结果，因为阴性结果是正确的。

某人处于哪一列取决于疾病的自然分布。某人在A列的几率（这个人实际上患有该疾病的几

率）取决于疾病的基础概率（base rate）。如果总人口中有5%的人患有该疾病，那么就有5%的人在A列中。

3.12.2　理解乳腺癌筛查

乳腺癌是可以进行诊断筛查检测的一种严重病情。乳腺癌筛查先从乳房X线检测开始。如果X线检测结果是阳性，则需要进一步检查：再一次做乳房X线检测、超声波或组织检查。

我们首先对回答乳腺癌筛查的敏感性和特异性的问题感兴趣。通过乳腺癌的基础概率信息，我们能回答最重要的问题：

> 如果一名女性获得一个阳性结果，她患有乳腺癌的可能性是多大？

通过咨询你的医生或做一些研究，你也许会发现乳房X线的敏感性大约是90%。特异性大约是92%。



因为有不同的人参加这项检测，乳腺癌筛查的准确敏感性和特异性随着时间推移而改变。相比过去，在年轻的女性中发现乳腺小肿瘤的现象更为常见。年轻女性的敏感性和特异性都不如年龄大的女性。当然，你应该和一名内科医生或专家核对目前的准确性水平。

表3-13按照表3-12的布局呈现了那些数据。因为A列和B列必须相互独立，总和为100%，我们同样能估计假阴性和假阳性的比率。

表3-13：10 000名女性的乳房X线检查的理论结果

	实际患乳腺癌的病人（A） N=120	实际未患乳腺癌的病人（B） N=9880
X光检查显示癌症	敏感性90%，N=108	假阳性 8%，N=790
X光检查未显示癌症	假阴性 10%，N=12	特异性92%，N=9090

表3-13同样基于总体中大约1.2%乳腺癌的基础比率，展示了10 000名女性的假设结果。



由于可以通过不同方式定义相关总体，所以很难确认乳腺癌的准确发病率，当然，还受乳腺癌检测结果准确性的限制。我使用的是目前针对40岁至84岁至性，经常报道和被广泛接受的患乳腺癌百分比的估计。

在解释医学检测结果前，我们先回到重要问题列表的第三个问题。如果一个人的检测结果为阳性，这个人患病的可能性是多少？10 000个进行乳腺癌筛查的女性中，有898人的结果为阳性。这898人中，790个人的结果是错误的，她们实际上并没有乳腺癌。898人中，有108个人的检测结

果是正确的，她们的确患有乳腺癌。换句话说，如果一个人的结果为阳性，那么她只有12%的可能性患病。对于定期进行X光检查，结果为阳性，最常见的结果是：病人实际上并无癌症。

那么阴性结果的准确性如何呢？在9102名筛查获得阴性结果的女性中，有12个人实际患有癌症。这是一个相对较小的数字，1%的1/10，但是检测会完全忽略掉这些人，他们不会得到治疗。

3.12.3　生效原理

托马斯·贝叶斯（Thomas Bayes）是18世纪的哲学家和数学家。医学筛查准确性，是使用了托马斯·贝叶斯条件概率的泛化方法的一个具体应用。“如果这样，那么……的几率是多少”，这是一个条件概率问题。

贝叶斯的条件概率方法是，看事件发生的自然概率。如果某人的检测结果为阳性，那么估算他患病几率的基本公式是：

$$\frac{\text{真阳性}}{\text{真阳性}+\text{假阳性}}$$

如果以条件概率来表述，公式如下：

$$\frac{\text{基础概率}\times\text{敏感性}}{(\text{基础概率}\times\text{敏感性})+(1-\text{基础概率})(1-\text{特异性})}$$

要回答我们乳腺癌例子中的重要问题（“如果一个女人的检测结果为阳性，她患乳腺癌的可能性是多少”），用乳房X光检查的公式套用这些值：

$$\frac{0.012\times0.90}{(0.012\times0.90)+(1-0.012)(1-0.92)}=0.1202$$

3.12.4　作出明智的决策

医学检测用来表明患者是否可能患病或处于即将患病的危险之中。识别疾病（比如癌症）是否存在的过程如下。通常至少有两个步骤，第一步对患者进行筛查检测，一般是相对简单和无创伤性的检测，用来寻找一个人可能患某种疾病的迹象。如果结果是阳性，则第二步进行第二次检测（或一系列的检测），这通常更复杂，具有创伤性，并且价格昂贵，而且也更加精确，以确认或驳斥原来的结果。

医学检测不是完全可靠和有效的。检测结果可能是错的。接受过医学检测的人有四种可能的结果。病人可能患病，并且检测也表明了这一点，或者病人没有患病，检测也没有发现疾病的存在。在这些情况下，检测起到作用并且分数是有效的。反之，检测结果可能反映了真实身体状况的相反情况，用一个阳性的结果错误地指示患有本不存在的疾病，或者用一个阴性的结果错误地

指示该患者没患病。在这种情况下，检测没起作用，其结果是无效的。这种结果的表格，类似于在统计决策中接受或拒绝假说的概率[Hack #4]。

当某人患有乳腺癌时，乳腺癌筛查非常容易发现这种疾。然而，这样一个针对低发病率的疾病敏感检测有一个缺点：更多的人将被告知她们可能得了这种病，但实际上她们并没有得。在医学检测的检测敏感性和检测特异性之间有一个折衷。更敏感检测往往会导致更多的假阳性，但在生死攸关的严重病情下，这似乎是一个我们能够接受的结果。

3.12.5　参阅

Gigerenzer, G. (2002). *Calculated risks. How to know when numbers deceive you*. New York: Simon and Schuster.

第 4 章

逆境制胜

（*Hack #35~#49*）

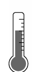
当你冒险时，为什么要承担不必要的风险呢？赌场游戏需要你冒点险，但本章的真实世界统计Hack将帮助你保持自身优势，也许甚至能帮你克服赌场优势。

我们先从得州扑克[Hack #36]（听说过吧）开始。然后是扑克[Hack #37]和概率游戏[Hack #38]。

当然，无论你玩什么，请确保总是明智地下注[Hack #35]。虽然当谈到你所冒风险的水平时，有些游戏[Hack #39和Hack #40]比其他游戏要好[Hack #41]。

如果你想和好友进行友好下注或与陌生人进行陌生下注，可以使用统计的威力来赢得一些令人惊讶的赌局，可以用扑克[Hack #42和Hack #44]、骰子[Hack #43]，或几乎任何你能想到的东西[Hack #46]，甚至包括你朋友的生日[Hack #45]。

说到怪异的赌博游戏（我认为我们也很怪异），玩这些游戏时，哪怕只是抛硬币[Hack #48]，你都需要知道一些奇怪的统计怪癖[Hack #47和Hack #49]。



HACK #35

4.1 明智地下注

不管是什么游戏，如果涉及金钱和几率，就有一些基本的赌博真相，而这可以帮助幸福的统计学家保持快乐。

虽然本章都是针对特定游戏的Hack，其中大多数游戏是几率游戏，但也有各种各样的对所有赌徒通用的技巧和工具。太多的神秘、迷信、数学方面的困惑充斥着赌博世界，了解赌博世界的地形可以帮助你行走自如。这个技巧通过教你下面的事情，来展示如何更明智地下注。

- 赌徒谬误，一个直观但虚假的信仰体系，除了见多识广的玩家，很多玩家都为此花了不少钱。
- 赌场和金钱。
- 系统、复杂的资金管理和无效的投注方法。

4.1.1 赌徒谬误

你是否有过这种经历：玩21点时，你连续抽到很多把差牌，由于你知道这种局面将会随时改变，所以你增加了赌注。如果是这样，那你就是屈服于*赌徒谬误*，它是这样一种信念：从长远来看，因为有某种预期概率，短期连续的坏运气可能会迅速改变。

赌徒谬误，是一个几率的游摆，它在坏结果的区域摆动一会儿，失去动力，然后摆回好结果的区域，在其中摆动一会儿。这种心态的问题是：和在靠运气的纯几率游戏中一样，运气是一系列的独立事件，每个人的结果和它之前的结果无关。换句话说，在好区域位置摆动或是在坏区域位置摆动和它前一秒的位置无关，这就是难点所在，甚至根本就没有摆动。变幻无常的命运手指在可能的结果之间随机弹取，并且出现任何结果的概率和每个结果相关。没有推动力（也看不出发展趋势）。这个真相经常被概括为“骰子没有记忆”。

与赌徒谬误信念一致的例子有下面这些：

- 一个一段时间内都没有吐子的老虎机要吐子了；
- 一名整夜坏手气的扑克玩家，很快就会得到一个超级大手，使局面逆转；
- 前3场比赛均失败的棒球队更容易赢得第四场比赛；
- 因为掷骰子时不太可能连续获得3个7，所以掷出3个7后，想马上再掷得第四个7基本是不可能的；
- 一个轮盘球已经连续8次落在红色数字上，下次几乎一定会落在黑色数字上。

如果能竭尽全力地避免这种谬误，那么赌博应该会让你少花一点钱。

4.1.2 赌场和金钱

赌场赚钱。赌场赚取利润的一个原因是，游戏本身赢得的金额比公平情况下赢得的金额略少。在一个几率游戏中，一个*公平*的彩金让双方参与者（赌场和玩家），从长远来看都收支平衡。

一个公平彩金的例子是，赌场使用只有36个数字的轮盘，一半红色一半黑色。命中一个红色数字后，赌场会将那些押注红色的玩家的赌注增加一倍。有一半的时间赌场会赢，有一半的时间玩家会赢。实际上，美国赌场使用38个数字，其中有两个既不是红色也不是黑色。这使赌场相对于公平的彩金有2/38的优势。当然，赌场通过这种方式赚钱，从一般意义上来说，这是公平的，

4

赌徒和赌场都希望这样，它是赌徒与赌场社会契约的一部分。不过，事实是，如果赌场仅依靠这个优势赚钱，那么没有人愿意继续经营。

赌场赚钱的第二个原因是，赌徒没有取之不尽的赌本，他们不会无休无止地赌下去。比如轮盘赌，赌场优势是5.26%，这是一个赌徒赌无数次，赌场能赚到的钱。这个赌无限次的赌徒一会儿赚钱，一会儿亏钱。在任何给定的时间，平均来说，他的赌金都将比开始时少5.26%。

虽然现实生活中真实情况是，大多数玩家通常会在用完筹码后的某个时间不玩了。大多数玩家有钱时继续投注，没钱时停止下注。当然，有些玩家赚钱时会选择离开。但是，没有玩家在没钱（没信用卡）时还下注。

试想一下，表4-1代表任何赌场游戏的1000名玩家。所有玩家一开始都有100美元，打算玩一晚上（4小时）。我们假设赌场有5.26%的优势，就像轮盘赌一样，虽然其他游戏有更高或更低的优势。

表4-1：1000名假想赌徒的命运

赌博时间	剩余赌金	平均剩余赌金	输掉的赌金	还在玩
玩1小时后	900	94.74美元	100	900
玩2小时后	800	94.74美元	200	800
玩3个小时后	700	94.74美元	300	700
玩4个小时后	600	94.74美元	400	600

在这个例子中，虽然使用的是虚构数据，但我敢打赌，这是保守的数据——4小时后，玩家仍然有56 844美元，赌场有43 156美元，从可用资金总量来看，赌场拿了43.16%。这比赌场官方的5.26%的优势要高。

玩家继续玩下去的倾向是人类行为，而不是和特定游戏相关联的概率，这种行为使赌场能够通过赌博获利。因为赌场的规则被发布和报道出来，所以统计学家可以为任何特定的游戏算出赌场优势。

但是，没人要求赌场报告他们从桌上游戏赢得的具体金额。但是，根据内华达（我最喜爱的赌场）劳克林（Laughlin）Lum旅游酒店里粗毛地毯的厚度，我猜测赌场赚得不少。这里一般赌徒的Hack是过一段时间后走开，不管你是领先还是落后。如果你在时间耗尽前足够幸运，能够遥遥领先，那么考虑走出赌场。

4.1.3　系统

基于资金管理和改变标准赌注数量，有几种不同的投注系统。一般系统建议在输掉一局后，增加赌注，当然也有些系统建议赢得一局后，增加赌注。由于所有这些系统假设连胜或连败，过

热或过冷，总是更可能终止而不是继续，所以多少有点基于赌徒谬误。但是，即使这样的系统具有数学意义，任何时间下注者必须增加投注，一直到玩家获胜，从长远看，有限口袋大小定律破坏了系统。

这里有一个真实的故事。我年轻时，第一次去一个合法赌博场所，我急于使用自己设计的系统。我注意到，如果我在轮盘赌中对12个数字下注，赔率是2：1。也就是说，如果我赌10美元而且赢了，除了能拿回10美元本金外，还能赢得额外的20美元。当然，这12个数字中任意一个出现的几率都不大，但如果我赌*两组*12个数字，那么我的胜算就大了。我有24/36（好吧，其实是38）的可能性会赢——几率超过50%！

当然，我明白，我无法通过投注两组数字赢得3倍的钱。毕竟，对于没有转到那12个数字的一组，我将失去一半的赌注。我明白如果我下注20美元，约2/3的时间我会赢回30美元。这将有10美元的利润。此外，如果转盘第一转我没有赢，我会再次赌上相同的号码，但这次我将赌注加倍！（我是一个超级天才，你同意吗？）如果我在第二转中也输了（可能性很小），我会再一次将我的赌注*加倍*，然后赢回我所有的钱，再加上50%的利润。总之，我就照计划做了，在三次转盘中都输了，所以没钱度过漫长的周末，也没钱坐22小时的车回家。

这种系统最简单的形式是在你每次赌输后将赌注加倍，然后当你赢（你一定会赢的）的时候，你就扳回了一点。问题是，连续输的情况很常见，这些都是几率的正常波动。在连输的过程中，不断将赌注加倍迅速用光了你的赌金。

表4-2给出了连输6轮、每输一次赌注加倍的结果，这种情况经常发生在21点、轮盘赌、骰子、视频扑克等游戏中。

表4-2：“赌输后赌注翻倍”系统

输的次数	赌注大小	总支出
1	5美元	5美元
2	10美元	15美元
3	20美元	35美元
4	40美元	75美元
5	80美元	155美元
6	160美元	315美元

连续输6次，即使是在输赢可能性几乎一样的游戏里，如赌轮盘赌的颜色，如果你玩的时间不只是一两个小时，这很可能发生在你身上。在一次试验中，赌输的实际几率是52.6%（20个输的结果除以38个可能结果）。对于任意连续6次转盘，玩家全部输掉的几率是2.11%（0.526 × 0.526 × 0.526 × 0.526 × 0.526 × 0.526）。

试想一下，两个小时内玩100转。玩家预计可以出现两次六连败。那么，一般情况，在此系统下，玩家被迫下注的金额是原始赌注的32倍，只为赢得等于原赌注的金额。当然，大部分时候（52.6%），出现连续六次失败后，就有连续第七次的失败。

帮助玩家在赌博游戏中作出明智战略决策的系统确实存在，如21点（算牌）和扑克（看穿你的对手），但在纯几率游戏中，统计学家们学会了接受能够预料到的情况。



4.2　知道何时持牌

在得州扑克中，“四规则”使用简单的计数来估计你赢得所有筹码的几率。

无限下注得州扑克无处不在。写这篇文章时，我把我的卫星天线指向ESPN、ESPN2、经典ESPN、福克斯体育、精彩电视台（Bravo）或是E频道，我能看到职业扑克玩家、幸运的业余爱好者、大名人、小名人甚至（主啊帮帮我们吧，在高速频道）NASCAR车手都在玩这个简单的游戏。

你可能自己玩这个游戏，或至少观看这个游戏。这个游戏最流行的版本很简单。所有玩家以相同的筹码开始。当他们的筹码都没有了时，他们也就离开了。每一轮中，玩家得到两张牌，这两张牌只有他们自己（和有专利的牌桌上的小相机）看得到。然后，将3张公共牌的牌面翻转朝上，这叫做*翻牌*。随后将另一张公共牌，牌面翻转朝上，这就是*转牌*。最后，再一张公共牌，*河牌*，将其牌面翻转朝上。每个阶段都有投注。玩家使用这7张牌中的任意5张（5张公共牌，加上他们手里的两张牌）来组成他们能组成的最好的5张牌。所有5张牌的组合中，最大的组合赢得比赛。

因为有些牌正面朝上，所以玩家能获得一些信息。他们也知道自己手里的牌是什么，这样他们掌握的信息就更多了。他们还知道一副标准的52张扑克牌中所有牌的分布。所有这些已知的关于值分布的信息[Hack #1]，使州扑克有很好的机会处处使用统计Hack[Hack #36和Hack #38]。

一个特别关键的决策点是：翻牌后那轮的下注。还有两张牌，可不可能提高你的手牌。如果你还没有最佳手牌（nuts），知道下两张牌有多大几率能提高你的手牌也是不错的。*四法则*使得你能够轻松且相当准确地估算这些几率。

4.2.1　工作原理

四法则工作原理如下：数出（数的时候不要动你的嘴唇）一副牌里能够帮你提高手牌的牌的数量，把这个数字乘以4。所得乘积就是你得到一张或多张这种牌的几率。

1. 示例1

你有一张方片J和方片3。翻牌是梅花K、方片6和方片10。你有4张牌冲击同花，有9张牌能使你获得同花。当然，其他牌也可以帮你（比如一张J会帮你组成一对J），但不是以让你觉得会赢的方式。

因此，有9张牌会真正帮到你。四法则估计你有36%的几率（$9 \times 4=36$）在转牌或河牌时达到同花。所以，你有大约1/3的几率。如果能在不投入太多筹码的情况下继续玩下去，那你也许应该通过跟注继续玩下去。

2. 示例2

你有一张方片A和一张梅花2。翻牌拿到了红心K、黑桃4和方片7。你可以算出，有6张牌能真正帮到你：3张A或3张2的任何一张。如果你赌到最后，一对2很可能只意味着麻烦，所以假设你希望看到的牌有3张，都是A。你只有12%的机会（$3 \times 4=12$）。弃牌吧。

4.2.2 生效原理

这里涉及的数学计算将一些重要的值进行了四舍五入，使得法则简化。思路如下：一副牌中大约还留有50张牌。（更准确地说，还有47张牌你没有见过）。当叫任意一张牌时，叫到你想要的牌的几率[Hack #3]是这个数除以50。



我知道，实际上是1/47。但我已经告诉了你，为了使得“四法则”容易记忆，一些东西已经被简化。

不管概率是多少，我们继续计算，因为你叫了两次牌，所以概率应该加倍。



这也不完全正确，因为在叫河牌时，牌池略小，所以你的几率会略高一点。

对于第一个例子，四法则估计同花的几率是36%。实际几率是35%。事实上，使用四法则的估计的和实际的几率往往相差正负几个百分点。

4.2.3 其他适用领域

注意，此方法也适用于只剩一张牌的情况，但在这种情况下，法则被称作二法则。将你想要的牌加和，然后乘以2，就能相当准确地估计只剩河牌时你获胜的几率。在大多数情况下，这个估计大约会偏离两个百分点，所以聪明的统计学扑克玩家称之为2+2法则。

4.2.4 不适用领域

随着能帮助你的牌的数量增加，四法则结果的偏离程度也会变大。当有12张出路牌（能帮你的牌）时，它是相当准确的，叫到这些有帮助的牌的实际几率是45%，四法则估计的是48%。但当有超过12张的牌可以帮助你时，四法则的估算结果会比实际高出不少。

为了不通过计算证实这一点，假设有25张牌（从47张牌中抽取）可以帮到你。这是一个绝佳场景（我至今无法想到会产生如此多的出路牌的场景），但四法则说你有100%的机会能拿到帮助牌中的一张。你知道这是不对的。毕竟，你叫的牌中有22张牌，完全不能帮到你。真正的几率是79%。当然，在这种情况下失算，不太可能伤害到你。在任何一种估计下，你都会一直赢到弃牌。



4.3　知道何时弃牌

在得州扑克中，底池赔率的概念提供了一个功能强大、决定何时跟牌何时弃牌的工具。

如果你在电视上观看扑克比赛，你会迅速学到一大堆行话。你会听到带A-K的成手（big slick）、一对A（bullets）、全押（all-in）以及输钱后不好的行为（tilt）等。你还将听到关于底池赔率的讨论，如："他或许会在这跟牌，不是因为他认为他有最好的牌，而是因为底池赔率。"

当底池赔率合适时，即使概率显示你会输，你也应该跟一手牌。那么，什么是底池赔率，为什么在我可能输的情况下，还要把更多的钱放到池子里？

4.3.1　底池赔率

底池赔率是通过比较你赢得底池的几率和赢得底池能获得的筹码量计算出来的。例如，如果你估计你将有50%的机会赢得底池，但底池太大，以至于赢得它能获得比跟牌成本两倍还多的钱，那么你应该跟牌。

看看到底池赔率是如何在实践中生效的，这里是一个有4名玩家的场景：西尔玛、路易丝、迈克和文斯。如表4-3所示，西尔玛在翻牌前处于最佳状态。



下面的表格显示了在一轮中的每个节点，每个玩家基于底池赔率作出的决策。从左至右，然后从上到下查看此表，看看西尔玛是怎么想的怎么做的，还有路易丝是怎么想怎么做的，等等。

表4-3：玩家起手牌

玩家	西尔玛	路易丝	迈克	文斯
起手牌	梅花A、红心2	梅花2、梅花4	红心4、黑桃5	方片K、方片10
开始的赌注	50	50	50	50

然后翻牌：黑桃A、方片3和方片6。表4-4是对每个玩家的处境分析。在翻牌后，他们中的3人都希望能改善他们的手牌，而他们中的一个——西尔玛，考虑到自己现在有最好的手牌，可能会满足现状，不打算改善手牌。西尔玛正驱动着下注，其他3名玩家决定是否跟牌。

表4-4：翻牌后的分析

玩家	西尔玛	路易丝	迈克	文斯
需要的牌		4个5里的任意一个	4个2或4个7里的任意一个	9张方片里的任意一张
获得牌的几率		16%	32%	36%
目前的底池	200	250	250	300
跟牌所花金钱占底池的百分比		20%	20%	17%
行动	下注50	弃牌	跟50	跟50

表4-4给出了翻牌后*底池赔率*的用法。西尔玛开始的时候有一对A，并在翻牌时获得第三个A。因此，她以下注开启每一轮。其他尚未获得任何好牌的玩家必须决定是否留下来，他们希望改善他们的手牌，让手牌变强，变成可能的赢家。

玩家需要作出跟牌或弃牌的决定时，底池赔率开始发挥作用。路易丝需要一个5来凑成顺子，她估计在下两张牌中有16%的几率能拿到一个5。然而，目前底池为250美元。其中西尔玛跟注了50美元，如果选择跟牌，路易丝将不得不支付底池的20%。这是20%的成本和16%赢取底池的几率对比。风险大于收益，因此路易丝弃牌。但是，迈克和文斯，有更多的出路牌，所以底池赔率建议他们继续玩下去。

接下来是转牌：梅花J。如表4-5所示，转牌后，只剩一张牌没发，迈克的底池赔率不再比他抽到一张赢牌更好，他弃牌。虽然和迈克相比，文斯开始时有潜在更好的手牌，当底池赔率表明他应该弃牌时，他也最终弃牌。

表4-5：转牌后分析

玩家	西尔玛	路易丝	迈克	文斯
需要的牌			和之前一样	和之前一样
获得牌的几率			18%	20%
目前底池	350		450	450
跟牌花费占底池的比例			22%	22%
动作	跟注100		弃牌	弃牌

我们假设玩家只使用底池赔率来做决策，不考虑他们很可能试图读懂其他玩家的影响（例如虚张声势、加注，等等）。顺便说一句，玩家使用四法则和2加2法则[Hack #36]计算他们获得一张能改善他们手牌的牌的几率。

4.3.2　生效原理

想象一下，有个游戏需要花1美元来玩。假设规则是这样的：一半的时间你会赢，并因此获

得3美元；另一半时间，你会输掉1美元并获得2美元。随着时间推移，如果你一直玩这个疯狂的游戏，你会获得一大笔钱。

在扑克中使用底池赔率，和这是相同的思想。有36%的几率促成同花，完全公平的下注是下底池36%的注。从长远来看，你会有36%的时间获得同花，达到收支平衡。如果你的支付少于底池的36%，长期来看仍然有36%的获胜几率，若你能在这样的游戏里玩牌，那你应该玩这个疯狂的游戏，对不对？好了，每一次你发现自己的处境是底池赔率比你必须下注的比例要高的时候，你就可以玩这样疯狂的游戏。相信统计学。玩这个疯狂的游戏。

4.3.3　其他适用领域

有经验的玩家不仅使用底池赔率对弃牌做决策，他们甚至用一个稍微更复杂的概念，叫做*隐含的底池赔率*（implied pot odds）。隐含的底池赔率不基于一个玩家必须跟注的数量占目前底池的比例，而是基于当那轮下注完成时，跟注占底池总数的比例。

如果玩家们仍然没有采取行动，一名犹豫不决是否基于底池赔率而留下的玩家可能期望其他玩家彻底跟进。这增加了最终底池的量，增加了如果他获得自己期望的牌所赢取的金额，并在所有下注完成后，增加了实际的底池赔率。

短语“隐含的底池赔率”有时也用来指：和所有下注轮数完成后最终总的底池相比，相对下注的花费。我也听说过这个词用来形容如果你碰巧“获得最佳手牌”（得到一个不太可能获得的强有力的手牌），或接近它的手牌，那么你很可能赢的比一般底池要多。有些玩家花费了大量的精力，进行了很多跟注，只是希望获得这些超级手牌中的一张，从而大捞一笔。

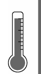
隐含的底池赔率是这样生效的。在表4-3这个场景中，迈克可能在转底后已经跟注了（第四张发的牌），他预计文斯也会跟。这将使最终的底池增加到650，使得迈克那轮的成本只有15%，并证明他跟注的正确性。

有趣的是，如果文斯投注时的底池已经含有迈克的跟注而变得稍大一点，那么文斯的100个筹码跟注的底池赔率将下降到18%，文斯可能会跟注。事实上，如果迈克是一个超级天才型玩家，他有可能在转牌时跟牌，他知道那样会改变文斯的底池赔率，因此这也鼓励了他跟注。现实生活中的职业扑克选手——真的，很好的职业扑克选手有时真是那么想的。

4.3.4　不适用领域

请记住，底池赔率基于这样的假设：你玩扑克的时间无限长。不过，如果在一个无限制的锦标赛中，由于你没有无限的资本，所以你可能不愿意基于从长远来看会发生什么的信念，而冒失去你全部筹码或大部分筹码的风险。

底池赔率基于的另一个生死攸关的假设是，你把“非常好的牌”视作能保证你会赢的牌。当然，事实并非如此。其他玩家可能也有非常好的牌，有比你更好的牌。


HACK #38

4.4 知道什么时候离开

在得州扑克中，当你“筹码短缺”时，你只有两个选择：立刻全押或过一会儿再全押。正如你可能已经猜到的那样，什么时候做最后一搏也是一个概率问题。

我在电视里听得州扑克锦标赛的扑克评论员谈论当筹码短缺时，下决定是如何如何“容易”。他们说它容易，是因为没有太多的选择。

“筹码短缺”这个术语可以有几种不同的用法。有时，它被用来指赌桌上拥有最少筹码的人。在这种用法下，即使你有成千上万的筹码并能付得起100底注和大盲注，如果其他人有更多的筹码，你也算是筹码短缺。

一个更好的、更适用于基于统计数据做决策的定义是：当你只能再付得起几次底注和盲注时，你就是筹码短缺。根据这一定义，赌上所有、希望能赢得两倍或三倍而回到游戏中的压力越来越大。我更喜欢这个用法，因为没有压力的话，“筹码短缺”的处境就没有太大意义。

但是当你筹码短缺，必须全押（赌上你的所有）时，这并不容易，容易吗？这非常非常难，原因有二。

- 你可能不会赢得比赛。你意识到你的筹码下降到了很少，不得不对下注进行好几次的加倍以回到游戏中。实际上，你怀疑你是否有很好的机会。这是令人沮丧的，在你悲伤时做任何决定都是困难的。
- 你犯了一个错误，你出局了。在这样高风险情况下，你没有多少犯错的余地，所以很难下决定。

运用一些基本的统计原则帮助决策，可能会使你感觉好点。至少你有一些不会感情用事的准则可以遵循。当你输掉时（你仍然可能会输，毕竟你处于筹码短缺状态），你可以怪我，或怪命运，不要怪自己。

4.4.1 辨识筹码短缺的情况

在比赛中，有时你的筹码非常少，以至于你将很快耗尽它们。除非你下注并很快获胜，不然你就会因盲注而用光筹码：强制下注的代价会把你的钱榨干。

究竟什么情况算筹码短缺？即使我们把筹码短缺定义为有多个大盲注（在一轮中，你被迫必须投注的两注里较大的那注），你需要多少这样的大盲注也因人而异，并没有一个统一的正确数字。这里有一些关于你面前有多少筹码就可以认为自己筹码短缺的不同观点。

1. 12倍的大盲注或更少

虽然你可以在不消耗完筹码的情况下再玩一段时间，但你会想在任意尚可的手牌上赌一把。你希望在这里赢得一些盲注。你赢得的盲注越多，你可用于等待杀手级手牌的时间就越长。如果别人对你加注了，你至少考虑以一个全押回应。

认为自己开始筹码短缺的玩家，希望在现在有好手牌的情况下全押，而不是在之后有普通手牌情况下被迫全押。开始冒险的另外一个优点是：公布“全押”后仍然会起一些作用。你将有足够的筹码让别人三思而后跟注。随后，你那可怜的小筹码将不足以摆布任何人。



当你采取全押希望导致对手弃牌时，要尽可能明智地选择你的对手。采取同样的全押策略，对手是小筹码时你的全押会比对手是巨大筹码时更有威力。同样的道理，如果你想跟注，面对拥有大量筹码的对手时，不要犹豫进行全押。他们会很乐意将你的赌注翻倍。

2. 8倍大盲注或更少

无论你在任何位置，庄家位、大盲注，还是先下注，在拥有任何前10的手牌时，考虑宣布全押。你依然有足够的筹码吓退一些玩家，尤其是那些拥有差不多等量筹码的人。

但是，你的筹码开始变得很少，少到你真的想被跟注。如果你可以低成本地玩一些低对牌（low pairs），试试吧，但如果你没有在翻牌中凑成三条（three of a kind），此时需摆脱困境。你需要保持尽可能多的大盲注，直到你有全押的机会。

下面是10手最有可能让你翻倍赢得筹码的牌：

- 一对A、K、Q、J或10；
- 同一花色的A~K、A~Q、A~J或K~Q；
- 不同花色的A~K。

3. 4倍大盲注或更少

这个时候，即使手牌有超过50%的几率会输，你也需要全押。故意下一注糟糕的赌注似乎有悖常理，但你正在和你希望翻倍却不断萎缩的基础筹码数做斗争。如果你等啊等，直到好的时机出现才全押，那么不管筹码还剩多少，你将不得不花好几倍额外的时间让自己回本。

底池赔率[Hack #37]在这时候开始生效。如果为了等待一个50%赢的几率而放弃25%赢的几率，那么你赢得的金额只有（如果）你有机会获得更好手牌时的一半。出现任何一对、一张A和别的什么牌、任何人头牌[1]和良好的起脚牌[2]，或同花连牌时，毫无疑问要全押。

注1：纸牌中的K、Q和J。——译者注

注2：在得州扑克里，2张起手牌中小的那张就叫做起脚牌。——译者注



当你的筹码非常非常少时（即，你的总筹码少于4倍大盲注），一个好的经验法则是：只要你拿到加起来是18或更好的牌，就全押。K算作13、Q算作12、J算作11，其余的牌是其面值。A算作14，但你已经在A与任意牌的组合下全押了，所以A算作什么无所谓。18点的手牌包括10~8、J~7、Q~6和K~5。

4.4.2 统计决策

当你宣布全押或者至少决定被套牢（如果被动情况下，有这么多筹码在底池，以至于你想全押）时，统计可以告诉你：在你输光所有筹码前，是不是有可能获得更好的手牌？”

我打算组50张看起来还不错、值得玩的得州扑克起手牌，这些牌能让你有机会赢取少数对手。我将使用3组，如表4-6至表4-8所示。虽然不同的扑克专家可能会对给定手牌的优良程度有争议，但筹码短缺时，大多数人都认为这些手牌至少可玩。



顺便说一句，每组中手牌不是按照质量排序的。

表4-6：10个很棒的起手牌

一　对	同一花色	不同花色
对A、对K、对Q、对J、对10	A~K、A~Q、A~J、K~Q	A~K

表4-7：15个不错的起手牌

一　对	同一花色	不同花色
对9、对8、对7	A~10、K~J、K~10、Q~J、Q~10、J~10、J~9、10~9、9~8	A~Q、A~J、K~Q

表4-8：25个还可以的起手牌

一　对	同一花色	不同花色
对6、对5	A~9、A~8、A~7、A~6、A~5、A~4、A~3、A~2、K~9、Q~9、10~8、9~7、8~7、8~6、7~6、6~5、5~4	A~10、K~J、Q~J、K~10、Q~10、J~10

当你筹码短缺时，盲注和强制性的底注都将来临，在你行动前，你知道自己有一定数目的剩余手牌。表4-9显示了你在下几次发牌中，得到很棒的、不错的或还可以的牌的可能性。

表4-9：获得可玩手牌的几率

手牌质量	下张手牌	5次发牌	10次发牌	15次发牌	20次发牌
很棒	4%	20%	36%	49%	59%
不错	7%	29%	50%	65%	75%

（续）

手牌质量	下张手牌	5次发牌	10次发牌	15次发牌	20次发牌
还可以	11%	46%	70%	84%	91%
还可以或更好	22%	72%	92%	98%	99%



我是这么计算表4-9的概率的：首先计算任何**特定**一对（你很可能同样得到一对A和一对2）的概率为0.004 5，然后计算任何两个**特定**的同花色但不同点数的牌的概率（0.003），再然后计算任意两张**特定**的不同花色且不同点数的概率（0.009）。下一步，对于每个类（很棒、不错或还可以的手牌），将那一类中的对数、不成对的同花手牌数以及不成对也不同花的手牌数分别乘以相应的概率，依此类推。最后，计算在给定机会数中没有获得期望手牌的概率，用1减去那个值，就得出表中每个单元格的值。

下面讲解如何使用表4-9。假设你筹码短缺时，刚刚发得了一手好牌。如果你认为在随后5手中的某个时刻你必须全押，只有20%的几率你会发到更好的牌。所以，你应该在这手好牌上全押。

如果你能再坚持20轮发牌，那么有大于50%的几率，你会得到能使你获得巨大成功的手牌，所以如果你想保险些，你现在可以不必全押。更常见的是，筹码短缺玩家甚至在没有排名前50位的手牌的情况下，考虑全押，例如不同花色的K~8这种。使用表4-9的概率，你可以放心地放下手牌，并希望在随后的5手中有更好的手牌。有72%的几率，你会得到这样的手牌。

最后，想象一下，你手里只剩下一些牌，因为盲注正让你的筹码趋于零。你往下看，看到一个像样的、还可以的手牌，如同一花色的8~7。表4-9能够使你回答一个重大问题：你的下一手可能会比这一手更好吗？你有11%的几率在下一手获得一个不错的或更好的手牌。所以，你的手牌不太可能得到改善。在这手牌上赌上你的未来。

4.4.3　理清思路

前面我们谈到，在筹码短缺时，为什么玩牌让人感觉如此困难。这里有一些帮你与进退两难做斗争的心理技巧。

- 现实一点

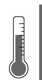
在21点牌中，闲家（player）抽到16，庄家（dealer）抽到7，闲家知道自己很可能会破21点。无论如何，他还是抽牌了，因为庄家可能有10点以下的牌，这给几乎不能赢的他最后一线生机。他知道他已经尽力给自己最好的幸存机会，这令他很高兴。同样的思路也适用于这里：你知道给了自己回到赌桌并赌赢的最好机会，这令你很高兴。

- 享受全押体验

没有什么比全押更激动人心。因为对于全押你没有选择的余地，所以放松下来，尽自己所能享受它吧。没有玩家会责备你做“这么愚蠢的事情”，因为你只是做了自己能做的最明智的事。

- 采取控制

为了避免使自己感觉被迫做了不想做的事，在不得不做之前就开始你卷土重来的计划。当你依然有10~12倍大盲注的筹码时，就要采取行动避免筹码短缺。这个时候的你比之后有更多的机会，所以你能够基于自身位置、对手以及马脚等，更精妙地发挥。你的筹码越少，你掌握自己命运的力量就越小。


HACK #39

4.5　在轮盘赌中输慢点

轮盘赌有很多漂亮颜色以及连小猫都喜欢的光泽。此外，你玩轮盘赌的时候看起来很酷。但是从长远来看，你会输钱，厌恶所有与之相关的事物。

像赌场中的大多数游戏一样，轮盘赌是一种纯几率的游戏。没有人能够预测小球最终会落在37（欧洲式）或38（美国式）区段的哪个区段。最好的玩家可以做的是知道概率、管理资金，以及假设自己会输掉。

当然，他可能是幸运的，可能会赢得一些钱，这最好不过了，但仍然会遵循大数法则[Hack #2]。从长远来看，如果他从未玩过这个游戏，他现在的钱很可能要多一些。事实上，如果他玩无限次，他一定会赔钱。（当然，大部分轮盘赌玩家没有玩无限次）要延长玩的时间，你应该知道和这个游戏相关的重要统计信息：转盘、轨道球、黑色和红色的布局。

4.5.1　基本赌注

图4-1显示了一个典型的轮盘游戏的投注布局。这是一个美式布局，这意味着有两个绿色的数字，0和00，（当小球落入这两个区段时，）不论你在红色和黑色或奇数和偶数上下注，赌场都不支付给你钱。欧式风格的轮盘只有一个绿色的数字0，相比美国赌场，减少了一半的赌场优势。

玩家可以以各种方式投注，这是轮盘在赌场如此受欢迎的原因一个。例如，玩家可以把一个筹码放在单个数字、两个数字、一种颜色甚至相邻的12个数字上，等等。和其他概率问题一样，随机获得期望结果的几率是预期结果（赢）的数目除以结果总数。

转盘上有38个间隔，由于所有38个可能结果的概率等同，所以计算是相当简单的。表4-10显示了玩家可下注的类型，赢得单次转盘和1美元赌注时赌场支付的实际金额、赌场优势等必要信息。

00	3	6	9	12	15	18	21	24	27	30	33	36	2 to 1
	2	5	8	11	14	17	20	23	26	29	32	35	2 to 1
0	1	4	7	10	13	16	19	22	25	28	31	34	2 to 1
	1st 12				2nd 12				3rd 12				
	1 to 18	EVEN							ODD	19 to 36			

图4-1：典型轮盘投注布局

表4-10：轮盘赌每1美元的统计数据

下注类型	获胜结果数	失败结果数	赔率	赌场支付	赌场优势公式	赌场优势
单个数字	1	37	37：1	35美元	（37–35）/（38/1）	5.26%
两个数字	2	36	36：2或18：1	17美元	（18–17）/（38/2）	5.26%
单色	18	20	20：18或1.11：1	1美元	（1.11–1）/（38/18）	5.26%
奇数或偶数	18	20	20：18或1.11：1	1美元	（1.11–1）/（38/18）	5.26%
12个数字	12	26	26：12或2.17：1	2美元	（2.17–2）/（38/12）	5.26%

赌场优势是这么计算的：首先确定如果没有赌场优势，赌场给每1美元的赌注回报是多少。公平的回报是：给赢家的金额等于所冒风险的量。所冒风险的量，实质上是可能输掉的结果数。然后，用没有赌场优势情况下应该支付给赢家的金额减去实际支付给赢家的金额。最后，将这些赌场持有的“额外”美元除以结果总数与获胜结果数的比值。如果没有额外的美元，游戏赌注在玩家间均匀地匹配，赌场优势是0%。

如果你研究表4-10中的轮盘统计量，一系列结论是显而易见的。首先，赌场通过假装轮盘赌上只有36个数字（即只有36个可能的结果）来赚钱，然后使用那个假装分布给赢家支付钱。

第二，不考虑轮盘赌的下注类型，赌场优势是常数5.26%。除了不知情的赌徒外，这是大家公认的，也是大多数赌场允许的。玩家往往允许在两个0和它们相邻的数字1、2和3，一共5个数字上下注。这是通过放一个筹码在0和1的交界线上完成的。你下这种赌注时，轮盘赌的掌盘人会检测确认，但是，在轮盘赌桌上这是最糟的赌法，且没有统计学家会这么建议。允许这样下注的赌场好像是在6个数字上支付。因此，赌场一贯的5.26%的优势在这里变成了7.89%，如表4-11所示。

表4-11：轮盘赌上5个数字的下注统计量（一个不明智的赌注）

下注类型	获胜结果数	失败结果数	赔率	赌场支付	赌场优势公式	赌场优势
5个数字	5	33	33：5或6.6：1	6美元	（6.6–6）/（38/5）	7.89%

4.5.2 生效原理

轮盘的流行，部分基于这样一个事实：有如此多不同类型的赌注。一个有很多筹码的赌徒可以将这些筹码在赌桌上散开，这样就在不同数字上或不同数字组合上下注。只要他在赌桌上避免最糟糕的赌注（5个数字），他可以放松地确信他的每次下注，赌场优势仍旧是5.26%。这是赌徒不用担心的。

事实上，单一的布局上可以有如此多的赌注种类，以至于不会有幸运的偶然事件。使用36个数字的决定是明智的，毫无疑问，它是多年以前制定的，因为有大量的因素将其指向数字36。当然，36不仅可以被1整除，也可被2、3、4、6、9、12和18整除，这使得许多简单的赌注成为可能。



4.6 在 21 点游戏中赢钱

或许对统计黑客最有潜力获利的应用是在21点牌桌上。

在21点游戏中，玩家的目的是让牌的总和比庄家的牌更接近21点（不超过21点）。这真的是一个简单的游戏。开始时你有两张牌，并可以尽可能多地要牌。人头牌的值10，A的值可以是1或11，其他牌的值是其面值。

如果你超过21，或者庄家比你更接近（但是不能超出），你就输了。下注输赢的机会均等，除非得到一个*黑杰克*：两张牌加起来和是21点。通常情况下，当你得到21点时，你获得3：2的赔率。庄家有一个优势，即他在你行动前不必采取行动。如果你爆了（超过21），他自动获胜。

统计学家可以通过使用两种来源的信息，来明智地玩这个游戏：庄家牌面朝上的牌和之前发的牌。基于概率的基本策略会让聪明的玩家不必太注意或学习复杂的系统，几乎就能和赌场平等对抗。考虑分析已发牌的方法统称为算牌，使用这些方法可以让玩家有统计优势。



美国法院裁定，在赌场里算牌是合法，虽然赌场希望你不要算牌。如果他们认为你在算牌，可能会让你离开这个游戏，去玩一些其他游戏，或者他们直接禁止你进入赌场。他们有权利这么做。

4.6.1 基本策略

先说一些重要内容。表4-12给出了依据你发到的两张牌和庄家牌面朝上的牌，21点牌合适的基本玩法。大多数的赌场允许你*分牌*（拿到一对牌，把它分成两幅单独的手牌）和*双倍下注*（将

你的赌注加倍，以换取再拿一张牌的机会）。是否应该停牌、拿牌、分牌或双倍下注，取决于你改进或损害手牌的可能性，以及庄家爆牌的可能性。

表4-12：针对庄家明牌的基本21点策略

你的手牌	拿牌	停牌	压双倍	分牌
5~8	总是			
9	2,7~A		3~6	
10~11	10或A		2~9	
12	2、3、7~A	4~6		
13~16	7~A	2~6		
17~20		总是		
2、2	8~A			
3、3	2、8~A		2~7	
4、4	2~5、7~A			6
5、5	10或A		2~9	
6、6	7~A			2~6
7、7	8~A			3~7
8、8				总是
9、9	2~6、8、9			7、10、A
10、10		总是		
A、A				总是
A、2	2~5、7~A		6	
A、3或A、4	2~4、7~A		5或6	
A、5	2或3、7~A		4~6	
A、6	2、7~A		3~6	
A、7	9~A	2、7~A	3~6	
A、8、9、10		总是		



在表4-12中，“你的手牌”指两张已经发给你的牌。例如“5~8”指你的两张牌加和为5、6、7或8。“A”表示王牌A。空白的单元格表示你不应该选择此项，或者在分牌的情况下，它甚至是不允许的。

其余的4列呈现的是典型的选项，告诉你庄家牌如何时，你应该选择的策略。正如你所看到的，对于大多数手牌，只有部分选项是具有统计意义的。此表显示了最佳方案，但并不是所有的赌场都允许你在有任何手牌时压双倍来分牌。但是，大多数赌场允许你拆分任何一对牌。

4.6.2 生效原理

和表4-12中的决策相关的概率，是由一些核心法则生成的：

- 庄家必须一直要牌，直到他达到17点或更高；
- 如果你爆牌了，那你就输了；
- 如果庄家爆牌了，但你没有爆，那你就赢了。

因此，主要策略是：如果庄家有可能爆，你自己就不要冒险。相反，如果庄家可能有很好的手牌，如20，你应该尝试提高你的手牌。能给你带来最大获胜几率的选项在表4-12里。



基于很多常用的、计算了某些结果发生概率的表格，我们给出了这些建议。表格里的统计数据或来自数学方法，或来自电脑模拟的数百万的21点数据。

这里有一个例子，从中可以看出当庄家的明牌是6点，我们是如何计算概率的。庄家的暗牌可能为10点，这实际上是最有可能的，因为人头牌计为10。如果庄家的暗牌是10点，那很好，因为如果庄家开始时是16点，他会爆的几率约为62%（如果你拿到16点，你也很可能会爆）。

由于有8张不同的牌会让16点爆掉（6、7、8、9、10、J、Q和K），所以爆牌的几率计算如下：

$$8/13=0.616$$

当然，最理想的结果就是庄家有一张10点的暗牌，但实际上庄家的暗牌不为10点的几率更大。暗牌是其他牌的可能性（9/13）大于10点牌（4/13）的几率。

除了A，任何牌都会导致庄家继续要牌。下一张牌会破坏庄家手牌的几率取决于庄家的实际起手牌。把它们全部加在一起，庄家的明牌是6点时，爆牌的几率不到62%。庄家的明牌为6点时，爆牌的实际几率接近42%，这意味着有他有58%几率不会爆牌。

现在，假设你有16点，庄家的暗牌是6点。你拿一张牌，爆牌的几率是62%。你立马输掉的几率是62%，庄家获得16点的几率是58%，将这两者进行对比。因为相比不要牌，要牌会输的几率更大（62大于58），你应该在庄家6点时停叫，如表4-12所示。

不同起手牌对庄家明牌的所有可能分支形成了表4-12的建议。

傻客投注（sucker bet）

如果庄家的明牌是A，许多赌场都提供机会让你买**保险**。保险意味着你下原始赌注一半的赌注，如果庄家有一个黑杰克（暗牌是10或人头牌），那么你就赢了边注，但是输掉了原始赌注（除非你也有一个黑杰克，这种情况下，你们打成平手，你可以拿回你的赌注）。

庄家有一张10点暗牌的几率是4/13，或31%。你输掉保险的几率要比你赢的几率更高。除非你算牌，从来不买保险。是的，即使你有一个黑杰克。

4.6.3　简单的算牌方法

本Hack之前描述的基本策略的前提是，你不知道牌堆里还剩什么牌。不管是使用单副牌、6副牌还是任意副牌，在特定游戏中前文假定牌的分布依然是原始分布。但是，发牌后，不管发出了什么牌，实际的概率都改变了，如果你知道新的概率，对于如何玩你的手牌也许会有不同的选择。

有一些精巧且合理（统计学上来说）的方法用于跟踪之前发的牌。如果你认真学习这些方法，让自己成为算牌员，你就有更多的机会和优势。但是，篇幅有限，我无法在这里提供一个完整而全面的系统。但对于我们这些愿意增加胜算的人，有一些无需特别努力或背诵许多图表就能提高胜算几率的方法可供利用。

提高获胜几率的基本方法是：当你有更好的获胜机会时，增加你的赌注。你必须在看到你的牌之前下注，所以当你的赔率有改善时，你需要提前知道。我们按照复杂顺序，分别讲解以下3种方法，让你知道什么时候增加赌注。

1. 算A

除非你被发到黑杰克，否则所有你赢的钱都是你下注的钱。当你有黑杰克时，你会得到一个3：2的支付（例如，每10美元的赌注赚15美元）。因此，当你获得黑杰克的几率比平均几率更大时，你可能要冒一个很大的险，下比平均更大的赌注。在其他条件相同的情况下，获得黑杰克的几率，是通过对两个概率求和得到：

- 先得到一张10点牌，然后得到一张A

$$(4/13) \times (4/51) = 0.0241$$

- 先得到一张A，然后得到一张10点牌

$$(1/13) \times (16/51) = 0.0241$$

把这两个概率加在一起，你会得到0.0482（约5%）的概率，这是你的起手牌就是21点的概率，也叫天生21点。

很显然，除非牌堆里有A，不然你无法获得黑杰克。当A发完后，你就没有机会获得黑杰克了。当有相对较少的A时，你获得黑杰克几率也就较小。一副牌中，如果先发出了一个A，你获得黑杰克的几率将降至0.0362（约3.6%）。如果发完1/4的牌还没出现A，你获得黑杰克的几率将提高到6.5%。



初露头角的算牌员要牢记：不要动你的嘴唇。

2. 算A和10点牌

当然，就像你需要一个A来获得黑杰克一样，你还需要有10点牌，如一张10、J、Q或K。当你在算A时，你也可以算发掉了多少张10点牌。

A和10点牌总共有20张，大约是总牌数的38%。当发完一半牌时，这20张牌应该也出现一半了。如果发出的这些关键牌少于10张，你得到黑杰克的几率会随之增加。发牌过半时，如果这20张牌都还没发，你得到黑杰克的几率会飙升至19.7%。

3. 通过点数系统算牌

当你玩牌时，你想要更多比例的大牌[3]（high card）和更少比例的小牌[4]（low card），一个简单的点数系统可以用来对一副或多副牌持续计牌。这比简单地算A或算A及10点牌，需要更多的脑力和专注力，但它提供了一个更加准确的指标——一副牌什么时候发这些有魔力的大牌。

表4-13显示了在这个点数系统下，一副牌中每张牌的点数值。

表4-13　简单的算牌点数系统

牌	点值
10、J、Q、K、A	–1
7、8、9	0
2、3、4、5、6	+1

一副新牌以0开始计数，因为这副牌中要发的–1点的牌和1点的牌数量相等。看到大牌是不好的，因为这意味着你得到黑杰克的几率有所下降，所以你的计数里失去了一个点数。发现小牌是好的，因为这意味着此时的牌里有更多比例的大牌，所以你获得一个点数。



你可以通过学习迅速识别普通对子牌的总点数，更高效和容易地学习算牌。一张大牌和一张小牌可以相互抵消，这样你就可以快速处理并忽略这类手牌。一对小–小的牌值为大点（2），一对大–大的牌很麻烦，这意味着每次看到这种令人失望的组合，你都要减去2点。

只有偶然情况下，你看到牌会使计数往好的方向大幅扭转。计数很少会远离0。例如，单副新牌中，前6张牌均是小牌的几率不到1%，而前10张牌均是小牌的几率约是1%的1/1000。

但是，计数不需要很高就能够将你获胜的几率提高到足以超越仅遵循基本策略时的水平。一副牌时，+2的计数就足够大到可以有效提高你获胜的几率。多于一副牌时，用你的计数除以牌的

注3：10、J、Q、K和A被称作大牌。——译者注

注4：2、3、4、5和6被称作小牌。——译者注

副数，这是对真正计数的良好估计。

即使只使用一副牌，有时你也会看到非常高的计数。当你看到那种一连串的幸运时，不要犹豫，提高你的赌注。如果你惯于使用点数系统，并详细了解这些系统，你甚至可以在要牌、停牌、分牌或压双倍时改变决策。

即使你只是使用这些简单的系统，也会在21点赌桌上提高赢钱的几率。但请记住，即使使用这些各种各样的系统，在赌场还有其他陷阱等着你，所以一定要始终遵循其他好的赌博建议[Hack #35]。



4.7 聪明地买彩票

你在大型彩票里中大奖的几率非常非常小，不管你怎么拆分这个大奖。但是，你确实对命运有一定的控制权。这里有一些方法能使你相比其他没有买这本书的彩票玩家更有优势（虽然优势是轻微的）。

2005年10月，最大的强力球彩票（powerball lottery）得主被加冕，并被授予3.4亿美元。那不是我。我不买彩票，因为作为一个统计学家，我知道，（相对于不买彩票的0中奖几率，）买只是稍微增加了我中奖的几率。这一点几率不值得我这么做。

当然，如果我不买，我就不可能中奖。买彩票不一定是坏的赌博，如果你要买，你可以做几件事以增加你赢钱（大概）的数量并提高你中奖（可能）的几率。在俄勒冈州的杰克逊维尔，10月那天，不管是谁买了那张能中奖3.4亿美元的彩票，他很可能都遵循了一部分制胜策略，而你也应该遵循这些制胜策略。

因为美国大多数州都有强力球彩票游戏，所以我们把它作为例子。但是，本Hack适用于任何大型彩票。

4.7.1 强力球赔率

像很多彩票一样，强力球要求玩家选择一组号码。然后抽取随机号码，如果你匹配部分或全部的号码，你就赢钱了！为赢得最大的奖金，你必须匹配很多号码。因为有这么多人玩彩票，所以售出了很多彩票，奖金也因此变得巨大。

当然，正确地选出所有中奖号码是很难做到的，但要赢得头奖，你就得正确地选出所有中奖号码。在强力球中，你先选择5个号码，然后选第六个号码：红色强力球。常规白色号码的范围是1~55，而强力球范围是1~42。表4-14显示了不同的中奖组合、奖金的数额，以及赢得奖金的几率和百分比。

表4-14：强力球的奖金

匹配	奖金	几率	百分比
只有强力球	3美元	1/69	1.4%
1个白球和强力球	4美元	1/127	0.8%
3个白球	7美元	1/291	0.3%
2个白球和强力球	7美元	1/745	0.1%
3个白球和强力球	100美元	1/11 927	0.008%
4个白球	100美元	1/14 254	0.007%
4个白球和强力球	10 000美元	1/584 432	0.000 2%
5个白球	200 000美元	1/3 563 609	0.000 03%
5个白球和强力球	特等奖	1/146 107 962	0.000 000 6%

7.4.2 强力球的奖金

像统计学家一样，用你现在可能有的所有智慧武装自己（除非这是你翻到的本书的第一个Hack），你可能已经对这个奖金一览表有了一些有趣发现。

1. 最容易的奖

只匹配强力球就能赢得最容易的奖，即使那样，获胜的希望也比较渺茫。如果匹配强力球（其他数字没有匹配），那你赢得了3美元。赢得这个奖的几率是1/69。

从任何理性的标准来看，这都不是一个很好的赌注。因为你花了一美元买一张彩票，玩一次，期望奖金是每69张彩票赢3美元。因此，平均而言，69次后你会赢得3美元，而你却花了69美元。

其实，你获得的奖金会比这多一点。表4-14所示的几率是基于一个特定的匹配，不考虑其他更好的匹配。当你匹配强力球时，部分时候你也会匹配一个白球，这时你的回报就是4美元，而不是3美元。选择5个白色球号码，至少匹配一个的几率是39%。

因此，匹配强力球之后，你有大于1/3的几率会至少匹配一个白球。即使这样，你的预期收益大约是每扔进老鼠洞69美元（我的意思是，花在彩票上）就收益3.39美元，这仍然不是一个好的赌注。

2. 只匹配强力球

只匹配强力球的几率似乎并不完全正确。我说过，强力球有42个不同的号码，所以概率匹配怎么会是1：69，不应该是1：42吗？

是的，但请记住表中展示的只是最糟情况下（没有匹配其他的球）获得的奖金。如果你把所有中奖可能组合在一起，你得到奖金的几率是1：37，3%左右。依然不是一个好的赌注。

3. 巨奖

巨奖的几率似乎也并不完全正确（好吧，好吧，我真的不希望你已经“注意到”了。我是直到做了一些计算后才注意到。）

如果从1~55号（白球）中抽出5个，从1~42号（红球）中抽出1个，然后做个快速计算，估计号码的可能性：

$$55 \times 55 \times 55 \times 55 \times 55 \times 42=21\ 137\ 943\ 750$$

换句话说，几率是1：21 137 943 750。或者，如果你想的更清楚一点，意识到随着号码被抽出，号码的总数变小，你可能会迅速计算可能的结果为：

$$55 \times 54 \times 53 \times 52 \times 51 \times 42=17\ 532\ 955\ 440$$

表格中显示的几率要比1：17 532 955 440好点。我第一次计算几率时，没有意识到号码顺序并不重要，所以任何剩余的号码可能会在任何时候出现。因此，下面才是正确的计算结果：

$$(5/55) \times (4/54) \times (3/53) \times (2/52) \times (1/51) \times (1/42)=1/146\ 107\ 962$$

4.7.3 赢得强力球

好了，（你可能认为）这些统计信息想要告诉我们，我们不应该再玩彩票了，因为，从统计学上来看，几率永远不会对我们有利。其实，用公平支付的标准来看，有一个时机可以玩，并要尽可能多地买彩票。

在强力球游戏中，当巨奖超过146 107 962美元时（或两倍的金额，如果你要一次性领取奖金），你应该买。只要它达到了146 107 963美元，买，买，买！因为从统计角度来看，匹配5个白球和1个红球的几率，正好是1比上那个大数字，当你的奖金比那个数字大的时候，它都是一个不错的赌注。

对于强力球和球的号码还有值的范围，146 107 962是一个神奇的数字。有观点认为，你获奖的几率并没有改变，但回报金额已增加至一个水平，在这个水平上，是值得买的，这类似扑克底池赔率的概念[Hack #37]。



你可以计算出任何彩票的“神奇数字”。一旦该彩票的回报高于这个数字，你就可以有理由地买一张彩票。使用我们例子中的“正确系列”的计算作为你的数学指导。问问自己有多少号码必须匹配，可能的号码范围是什么。请记住，每次你抽出一个球或号码时，就将你要除的那个数减1，除非号码可以重复。如果号码可以重复，那么在你的连乘中分母保持不变。

至于何时购买彩票，和计算的实际魔法数字、促使你无节制地购买彩票的奖金数量有关。宣传的所谓头奖金额，其实并不是头奖。宣传的“头奖”其实是若干年里获奖者领取的一系列部分奖金的总额。在赌博和统计意义上，你确认的金额，即真正的头奖是如果你选择一次性奖金会赢得的金额。一次性奖金通常比宣传的头奖金额的一半还少一点。

所以，如果你已经确定，你要买的彩票已经增长到了头奖金额，现在在统计学上是个购买好时机，那么你应该买多少彩票？为什么不每一个都买？为什么不花146 107 962美元买每个可能的组合？保证你会中奖。如果头奖大于这个数额，那么你会赢钱，一定的，对吧？嗯，其实不是。否则，我会很有钱，我也绝不会跟你分享这个技巧。为什么不能保证你赢？可能的情况是你会被迫拆分奖金！请看下节……

4.7.4 不要拆分奖金

如果你的彩票的确中奖了，那你希望自己是唯一的赢家，所以除了决定何时买，还有各种的策略来提高这个中奖号码仅归你所有的可能性。

首先，我们假设：中奖号码是随机选取的。我不想成为一个阴谋论者，我也不相信上帝有时间或愿意影响彩票中奖号码的抽取，所以我不会列出任何只在非随机抽取中奖号码下生效的战略。在你考虑如何选择彩票号码时，这里有一些合理的提示。

- 电脑选号

让电脑选，或者，至少自己选择随机数。随机数对其他玩家具有意义的可能较小，所以他们不太可能将随机数字选作他们自己的彩票号码。强力球的人报告说，所有中奖彩票中70%是由店内电脑选出的。（他们还指出，在“我们告诉你，结果是随机的”有点异想天开的想法中，所有购买的彩票70%是由计算机产生的）。

- 不选日期

不选可能是日期的数字。如果可能，避免小于32的数字。很多玩家总是选重要的日期，如生日、纪念日以及出狱日期，等等。如果你的中奖号码是别人的幸运日，就增加了你将不得不分割你奖金的几率。

- 远离知名号码

不要挑那些众所周知的数字。在2005年10月的强力球结果中，数百名玩家选择在热映的科幻电视剧《迷失》中起重大作用的数字，作为他们的彩票号码。这些人没能荣获大奖，但如果他们赢得了大奖，他们将不得不把百万奖金分割成数百片。



还有一系列纯粹的哲学技巧，和因果的抽象理论以及现实的本质有关。例如，有些哲学家会建议选上周的中奖号码。因为，虽然你可能不确定知道什么是真实的，在这个世界上什么能发生，什么不能发生，但你至少知道，上周的中奖号码是有可能成为本周中奖号码的。它之前发生过，它可能会再次发生。

虽然你赢得大奖的几率微乎其微，但你可以遵循一些统计学原理，做一些事情来真正掌握自己的命运。（顺便说一句，在意大利语中命运的单词是乐透lotto。）哦，还有一件事：在开奖日当天买彩票。如果你购买彩票的时间踞获奖号码公布时间还有很长，相比你赢得头奖的几率，你有更大的几率会被雷电击中、在浴缸中溺水或被一辆小货车撞到。时间就是一切，我不希望你错过机会。



4.8　好运玩牌

弗兰克叔叔花费很多时间在酒馆里玩骰子赢愚笨的酒吧注，还有冲女士们发出迷人的微笑，虽然这是事实，但他的生活比这丰富。比如，有时候他玩扑克，不玩骰子。

人们往往对不同组合牌出现可能性的理解水平自我感觉良好，尤其是卡牌玩家和扑克玩家。他们的经验已经告诉他们，一对、三条及同花等很少出现。但是，将这种直观知识运用到本游戏情境之外的其他卡牌问题是困难的。

我那精于统计的弗兰克叔叔知道这一点。有时候，弗兰克叔叔，我很抱歉地说，用他的统计知识作恶，不作善，他已经赢得了一批使用扑克牌的酒吧投注，并声称这帮他支付了研究生学费。我在此与大家分享，目的只是为了证明某些基本统计原则。我相信你会用新学到的知识取悦别人，打击犯罪，或赢得廉价的非酒精饮料。

4.8.1　获得小同花

在扑克中，同花是5张花色相同的牌。不过，对于我的叔叔弗兰克，不管他在什么地方，在他被要求离开前，很少有时间能发完所有的手牌。因此，弗兰克叔叔常基于他所谓的小同花（li'l flushes）下注。

1. 投注

一个小的同花（哎呀，对不起，我的意思是小同花）是任意两张相同花色的牌。弗兰克喜欢打一个赌，几乎总能赢，那就是在你手牌里发现两张相同花色的手牌。此外，由于时间限制，他的扑克手牌只有4张，而不是5张。

赌注是，你从一堆随机牌里发给我4张牌，我会得到至少两张同一花色的牌。虽然这似乎不

太可能，但实际上4张牌花色都不同的几率更小。我算过，一手4张牌，花色不同的几率大约是11%。所以，得到一个小同花的几率大约是89%！

2. 生效原理

有不同的方式计算扑克牌手牌的概率。对于这个酒吧赌注，我用的方法是这样的：数出可能获胜手牌组合的数量，并和所有手牌组合的总数相比较。这是在“好运玩骰子”里使用[Hack #43]的方法。

为了计算4张牌代表4种不同花色，即它们之间没有两张牌同花的几率，我们先计算出可能的4张牌的组合数量。试想一下，任何第一张牌（有52种可能），这张牌与任何剩余的第二张牌（52×51），加上第三张牌（52×51×50）和第四张牌（52×51×50×49），你会得到共计6 497 400种4张手牌的不同组合。

接下来，想象4张手牌的前两张牌。它们花色相同的几率只有0.235 2（51张牌中依然还有12张是同一花色）。因此，在所有可能的4张手牌组合里，大约有150万的组合会在头两张牌里形成同花。它们不是同花的几率是0.764 8。这使得头两张牌是不同花色的可能数量是4 968 601。

这一数量中，有多少不会收到和前两张牌花色都不同的第三张牌？还剩余50张牌，50张牌中有26张有还没出现过的花色。所以，第三张牌和前两张花色都不相同的几率是26/50（52%）。

这使得前3张牌花色都不相同的组合数量变成2 583 673。现在，这个数字里，有多少会抽出第四张牌是第四种没出现的花色？剩下的49张牌中有13张代表了最终的第四种花色。剩下的手牌中26.53%的牌将和第四张牌花色一样，4种不同花色的组合数量达到了685 464。685 464除以可能的手牌组合总数，结果是0.105 5（685 464/6 497 400）。

4张手牌有4种不同花色的几率是11%。哎呀！顺便说一下，一些超级天才可以只使用相关比例，就能得到相同的结果，这也是我们在不同的计数阶段一直用的方法，根本不必计算：

$$0.764\ 8 \times 0.52 \times 0.365\ 3 = 0.105\ 5$$

4.8.2 寻找两副牌的匹配

你有一副扑克牌，我也有一副扑克牌。这两副牌都洗过了。如果我们一次把它们发完，即一次性将两副牌都发完，它们会有匹配吗？我的意思是，他们会完全匹配吗？相同的牌，例如，我们两个都同时出现方片J？

1. 投注

大多数人会说没有，或者至少它会偶尔发生，但一定不会太频繁。令人吃惊的是，当你发两副牌时，你会频繁发现至少一对，而不频繁发现倒是不寻常的。如果你进行这样的下注或进行很多次

4

这项实验，在大多数情况下你将获得至少一对匹配。事实上，只有36.4%的几率你找不到一对匹配！

2. 生效原理

以下讲解如何从统计角度思考这个问题。因为牌被洗过，所以可以假定任何两张被翻转的牌代表的都是从两幅牌这个理论总体中抽取的随机样本。对于任何给定的一对牌，可以算出这对牌匹配的概率。因为你抽样52次，在这些抽样中得到匹配的几率会随着抽样次数的增加而提高。就像扔一对骰子得到7一样：在任意给定的一次投掷下，它不太可能，但随着投掷次数的增多，它变得非常可能了。

为了计算一系列结果中，命中某人希望结果的概率，我们先计算尝试若干次都没有得到希望结果的概率，这样数学运算会简单一点。对于任何给定的牌，该牌在另一副牌完全配对的几率是1∶52。不配对的几率是51∶52，或0.980 8。

但是，你不止一次尝试配对，你尝试了52次。那么，52次尝试都没得到一对匹配的概率是0.980 8的52次方。用数学语言表达就是0.9808^{52}。

等一秒钟，我会在脑中计算（0.980 8乘以0.980 8乘以0.980 8，以此类推，52次结果约为0.364 3）。好吧，所以它不会发生的几率是0.364 3。为了得到它发生的几率，我们用1减去这个数字，得到0.635 7。

你会发现在两副牌中，约2/3的几率会至少有一对匹配！非常好。去赢免费的柠檬水吧。



4.9　玩骰子行大运

下面是一些用诚实的骰子进行的诚实赌注。但是，这只是说明你没有作弊，并不意味着你不会赢。

人们对统计学家有一个不幸的刻板印象：戴眼镜的内向书呆子，永远不会和大家一起喝啤酒。这是如此荒谬的想法，以至于上周六、日，我在《龙与地下城》的每周例行聚会中，只想到这点，就大笑不止，笑得我的单片眼镜差点掉进雪利酒里。

事实是，在酒吧里展示简单的概率知识，会让顾客感觉非常有趣，会使你成为聚会的焦点。至少，对于我叔叔弗兰克而言确实是这样的，他多年来一直用他的统计技巧来赢得免费的饮料和腌蛋（或任何那种在大缸里的东西，总是在电视上能看到它们在酒吧里）。

这里有一些使用任何一对公平的骰子都能赢得赌注的方法。

4.9.1　骰子的结果分布

首先，让我们熟悉投掷两个骰子可能出现的结果。你应该知道，大多数骰子有6个面（我把它幻想成进行角色扮演的朋友，并称作*六面骰子*），6个面分别代表1~6。

计算可能的结果不过是列出这些结果，并计算它们。图4-2显示了掷两个骰子的所有可能结果。

掷出2有1种方式	掷出3有2种方式	掷出4有3种方式	掷出5有4种方式	掷出6有5种方式	掷出7有6种方式	掷出8有5种方式	掷出9有4种方式	掷出10有3种方式	掷出11有2种方式	掷出12有1种方式
1 1	1 2	1 3	1 4	1 5	1 6	2 6	3 6	4 6	5 6	6 6
	2 1	2 2	2 3	2 4	2 5	3 5	4 5	5 5	6 5	
		3 1	3 2	3 3	3 4	4 4	5 4	6 4		
			4 1	4 2	4 3	5 3	6 3			
				5 1	5 2	6 2				
					6 1					

图4-2：掷两个骰子可能的结果

这种分布形成了表4-15所示的频率。

表4-15：投掷两个骰子结果的频率

总投掷数	组合数	频　率
2	1	2.8%
3	2	5.6%
4	3	8.3%
5	4	11.1%
6	5	13.9%
7	6	16.7%
8	5	13.9%
9	4	11.1%
10	3	8.3%
11	2	5.6%
12	1	2.8%
可能结果的总数	36	100%

当然，掷骰子的游戏完全基于这些期望频率（expected frequency）。当你看这个频率分布时，可能会想到一些有趣的赌注。例如，虽然7点是最常见的投掷结果，很多人也都知道这一点，但它只是比6或8稍微高一点。

事实上，如果你不需要对某个具体数字下注，你可以赌在出现一个7之前会出现一个6或8。这些骰子完成投掷后，所有两个骰子之和中，有超过1/4的几率（约28%）会是6或8。这实质上比和为7的可能性要大，出现7的几率只有1/6。

4.9.2 用骰子进行酒吧投注

我的叔叔弗兰克以前经常和一些迟钝的顾客打赌，说在顾客掷得一个7以前，会掷得一个5

或9。14次打赌中，弗兰克叔叔会赢得8次。

有时候，老弗兰克打赌说，将一对骰子投掷一次，会出现6或1。不过，人们首先想到的是，似乎此情况发生的几率至少低于50%，事实上，出现一个1或6的几率约为56%。顺便说一下，任何两个不同的号码出现的概率与之相同，所以你可以用一个具有吸引力的陌生人的生日来挑选数字，也许因此开启你们之间的对话，使得你们结婚，育有子女，或两者兼而有之。

如果你比我叔叔弗兰克更诚实（你有98%的可能性比他诚实），这里有一些输赢参半的骰子投注。A列结果和B列结果发生的可能性相同：

A	B
2或12	3
2、3或4	7
5、6或7	8、9、10、11或12

对任何一种结果，胜负的几率是相同的。

4.9.3 生效原理

对于这个Hack展示的赌注，下面是计算的获胜概率：

赌　注	获胜结果数量	计算	结果比例
5或9 vs 7	8 vs 6	8/14	0.571
出现1或6	20	20/36	0.556
2或12 vs 3	2 vs 2	2/4	0.500
2、3或4 vs 7	6 vs 6	6/12	0.500
5、6或7 vs 8或更高	15 vs 15	15/30	0.500

“赌注”列表示两个有竞争关系的结果。（例如，得到7之前会得到一个5或9吗？）“获胜结果数量”列展示两种情况下不同投掷结果的数量（例如，8次获得一个5或9，6次获得一个7）。“结果比例“栏表示你获胜的几率。

你可以通过这些各种各样的赌注，用两种不同的方式获胜。如果是胜负几率均等的下注，从长远来看，你可以通过比对手少下注来赚取利润。他不会知道胜负几率是均等的。但是，如果几率有利于你，你就要考虑给你的目标提供稍微好点的回报，或选择很可能更频繁出现的结果。



4.10 提高卡牌的杀伤力

在得州扑克和其他扑克游戏中，有几个初步的技巧以及关于概率的一点基本知识，会立刻把你从绝对初学者推动到更高超的水平，甚至使你被当做作弊赌徒而陷入麻烦。

在一些重要方面，出现在电视上的专业得州扑克玩家跟我们有所不同。（好吧，他们可能和你只在几个重要方面有所不同；他们和我在很多重要方面都不同，以至于即使我的大脑像电脑一样也无法达到那么高的水平。）下面是他们玩扑克时已经掌握的两种技能：

- 知道在不同阶段（翻牌、河牌，等等）获得他们想要的牌的粗略概率；
- 快速识别也许被其他玩家持有的可能更好的手牌。

本Hack介绍一些技巧和工具可以助你从新手转变为半职业玩家。这些都是一些简单好用的知识和帮你快速作出决策的经验规则。就像这本书中的其他扑克Hack一样，它们提供的策略技巧完全基于统计概率，即假定一副标准扑克的52张牌是随机分布的。

4.10.1 改善你的手牌

在得州扑克中，你有一半的时间会得到对子或更好的手牌。我会重复一遍，因为这对理解游戏很重要。一半的时间（实际略低于52%），如果你留在赌桌旁的时间足够长，7张牌（你的两张牌，加上所有5张公共牌）中你将至少有一个对子。它可能一直在你的手牌里（称为*口袋对子*[5]或*连接对子*[6]），也可能由手牌中的一张和公共牌中一张组成，或者全都来自在大家都可以叫的公共牌里。

如果多数时候，每个玩家发到7张牌，平均每个玩家都会有一对，那么有一对低对（low pair）却要坚持到结束的你可能会输，当然，这只是从统计学上来讲。换句话说，另一个玩家至少有一对的几率大于50%，并且这一对可能是对8或更高的对子（13对中只有6对是对7或更低的对子。）

对子的常见性解释了为什么A被高度重视。很多时候，翻牌（head-up）战斗归根结底是对子和对子间的较量。另一个很好的时间比例是，A作为起脚牌或在决胜局中起着重要的作用。有A是好事，这一切都因为概率。

4

1. 概率

如果你知道一些常见期望结果的常见概率，就可以更明智地作出这样的决定：在试图降低对手数量时应该停牌还是加注。表4-16列出了抽到的那张牌能在不同阶段提高你的手牌的概率。这个概率根据一副牌里还剩多少牌、多少不同的牌可帮你（你的*出路牌*），以及还会发多少牌计算而得。例如，如果你有一个A~K，希望配对，有6张牌可以配对，换句话说，你有6张出路牌。如果你只有一张大牌A，但希望能找到另外一张A，那你就有3张出路牌。如果你有一对口袋对子，并希望在公共牌里找到强大的第三张牌，你只有两张出路牌。

注5：在得州扑克中，口袋对子（English Pocket Pair）是由玩家的两张底牌构成的对子。——译者注

注6：连接对子（wired pair）是由一个玩家的第一张和第二张牌组成的对子。——译者注

表4-16：改善手牌的概率

剩余发牌数	6张出路牌	3张出路牌	2张出路牌
5（翻牌前）	49%	28%	19%
2（翻牌后）	24%	12%	8%
1（转牌后）	13%	7%	4%

这里描述的情况假设你发到了两张牌。毕竟，在大多数扑克游戏中，在翻牌前下注是预计好的，不需要做任何决定。顺便说一下，因为你希望通过翻牌改善自己手中未成气候的牌，所以你想知道翻牌可能改善手牌的几率。它们分别是：

剩余出路牌	在翻牌中得到获胜牌的几率
6	32%
3	17%
2	12%

2. 启示

根据表4-16中所述的分布，这里有几个你应该牢记心中的简单的观察和启示。

一半的时候，你会配对。这对大牌来说是正确的，如A~K或小牌，如2~7。你甚至可以选择对已有的两张牌进行配对，它们配对的几率是28%。启示：在锦标赛中，当低筹码时，只要你获得A，就立马全押。

如果你没获得第三张牌，你需要在翻牌时把对子变成三条（3张一样的牌），并且你只有8%的几率能获得三条。启示：不要花太多的钱等你的低对子牌变成能使你获得巨大成功的手牌。

随着越来越多的牌被发出，你那翻牌前看起来还不错的A~K或K~Q如果还没有配对或凑成顺子，那它们的潜在优势会削弱。如果你在河牌时并没有命中，100次中有87次，那个伟大的起手牌仍然是为数不多的优质手牌。启示：只有当你可以花少量代价地这样做时，保持你那A~K未完成的梦想。

4.10.2　快速解读公共牌

这里有一些关于你对手手牌的常识，这些常识一定是真实的，很多人知道但不一定会说出来：

如果公共牌里没有……	你的对手不会有……
一对	四条
一对	满堂红
同一花色的3张牌	同花
5张顺子牌中的3张牌	顺子

你可以通过学习这些规则，分析对手可能有的手牌，从而更快做决策。那么，当某些情况不可能时，你可以自动排除杀手牌。你也许不必担心速度，但如果你无需浪费精力每次都从头开始搞清楚这些东西，你就可以花时间专注于更重要的决定。



4.11　让你最亲密的 23 个朋友震惊

一组人中，至少有两个人是同一天生日的几率是多少？根据现在的人数来看，这个几率出奇地高。使用简单的概率规则，可以在聚会上使你的朋友对你印象深刻（也许还能在酒吧打赌赢得一些钱）。

有些在逻辑上似乎不可能的事件，其实在某些情况下完全有可能。例如确定一组人中至少有两人的生日是同一天的概率。许多人震惊地得知，只要群组的人数不少于23，那么至少有两人生日相同的几率比50%还高！通过使用一些简单的概率规则，你可以算出任何规模的群组中，这一事件发生的几率，然后当你的预言成真时，你的朋友会吃惊。



你也可以利用这个结果在酒吧下注，从而赚一些钱（只要那里至少有23个人）。

那么，你如何算出至少两个人的生日是同一天的概率？为了解决这个问题，你需要对生日在总体中的分布做几个假设，并知道计算概率的一些规则方法。

4.11.1　入门

要确定至少有两个人生日是同一天的几率，我们必须对生日的分布做一些合理的假设。首先，我们假设生日在总体中是均匀分布的。这意味着在一年中，每一天出生的人数大致相同。

这一假设不一定完全正确，但非常接近真实情况，足以让我们相信计算的结果。然而，这个假设对2月29日这个日期是绝对不正确的，因为它只在每4年一次的闰年发生。好消息是，没有那么多的人出生在2月29日，所以我们能够在忽略它的情况下仍然得到准确的估计。

一旦我们做好了这两个假设，就可以相对容易地解决生日问题。

4.11.2　运用全概率法

在我们的问题中，只有两种*互斥*的可能结果：

- 至少有两人的生日是同一天；
- 没有人的生日是同一天。

由于这两件事情必有其一发生，所以其概率之和始终等于1。统计学家把这称作*全概率法则*，

而且在这个问题上派上了用场。



术语“互斥”意味着，如果一件事情发生，另一件事情就不会发生，反之亦然。

一个简单的抛硬币的例子可以帮助我们理解它的原理。抛一枚正常的硬币，得到正面的概率为0.5，得到反面的概率也是0.5（这是互斥事件的典型例子，因为抛掷硬币一次不能同时得到正面和反面！）。只要抛出硬币，两件事情必有其一要发生。它落地时一定要么正面朝上，要么反面朝上，所以正面或反面发生的概率是1（0.5+0.5）。进而我们能想到，正面的概率是1减去反面（1–0.5=0.5）的概率，反之亦然。

有时候，计算一件事件没有发生的概率很容易，所以可用该信息来确定它发生的概率。所有人生日都不同的概率比较容易弄清楚，它只取决于组里有多少人。

试想一下，我们这个组只有两个人。他们同一天生日的概率是多少？嗯，他们生日不是同一天的概率很容易计算：第一个人的生日是某天，第二个人的生日如果在其他的364天中的一天，那么他们的生日就不是同一天。所以，在数学上，概率是364除以365（可能的生日的总数），或0.997。

由于两个人生日不是同一天的概率是0.997（非常高的概率），实际上他们同一天生日的概率等于1–0.997（0.003，非常低的概率）。这意味着，每1000对随机选定的人，只有3对的生日是同一天。到目前为止，这从逻辑意义上来说是完美的。然而，一旦我们开始在群里添加更多的人，事情就开始改变（迅速改变)!

4.11.3　计算独立事件的概率

解决我们的问题还需要另一个诀窍，即采用独立事件的概念。如果两件事情同时发生的概率等于它们各自独立发生概率的乘积，那这两件事被视作*独立事件*。

我们再一次以典型的、简单而又易于理解的抛硬币为例。如果你抛两次硬币，两次都得到正面的概率等于正面的概率乘以正面的概率（0.5 × 0.5=0.25），因为抛一次硬币的结果对其他次抛硬币的结果没有影响（因此，它们是*独立事件*）。

所以，当你抛两次硬币，有1/4的概率会出现连续两个正面。如果你想知道连续抛出3个正面的概率，答案是0.125（0.5 × 0.5 × 0.5），这意味着连续3个正面发生的概率只有1/8。

在我们的生日问题里，每次添加一个人到组里，就相当于添加了一个独立的事件（因为一个人的生日不影响任何其他人的生日），因此，不管有多少人，我们都可以算出至少两个人同一天生日的概率，我们只需不断将概率相乘即可。

检查一下，不管我们的小组有多少人，只有两个相互独立的事件发生：至少有两个人生日是同一天或者没有任何两个人生日相同。由全概率法则，我们得知，我们可以计算没有任何两个人同一天生日的概率，然后用1减去这个概率就等于至少有两个人生日相同的概率。最后，我们也知道，每个人的生日都独立于组里的其他成员。都明白了吗？好，那我们继续！

4.11.4　解决生日问题

我们已经确定了两人小组中，两人生日不是同一天的概率等于0.997。假如我们在这个组里添加了另一个人，所有人生日都不同的概率是多少？对于第三个人来说，如果生日在其他的363天，那么他们3人的生日不同。因此，第三个人和其他两人生日不同的概率是363/365，或0.995（略低）。

但是，请记住，我们感兴趣的是，所有人生日都不同的概率，所以我们使用独立事件的法则，将前两个人生日不是同一天的概率，乘以第三人与这两个人生日不是同一天的概率：0.997 × 0.995=0.992。所以，在这个3人组里，所有人生日都不同的概率是0.992，这意味着至少有两个人生日相同的概率是0.008（1–0.992）。

这意味着，随机选择出1000组3人小组，只有8组会有至少2人生日相同。这仍然是一个非常小的几率，但注意，相比两人小组，3人小组的概率翻番了（从0.003变成0.008）!

一旦我们开始把越来越多的人添加到组里，至少2人生日相同的概率也随之增加得非常快。当我们的组员达到10人时，至少2人生日相同的概率高达0.117。我们应该如何确定这个值呢？对每个添加到该组的人，将他带来的额外分数和以往的分数相乘。每个额外分数都以365为分母，分子是365减去添加这个人之前小组的人数。

因此，对于我们前面提到的10人小组，最后分数的分子是356（365–9），概率计算如下：

$$\frac{364}{365}\times\frac{363}{365}\times\frac{362}{365}\times\frac{361}{365}\times\frac{360}{365}\times\frac{359}{365}\times\frac{358}{365}\times\frac{357}{365}\times\frac{356}{365}=0.883$$

这告诉我们，在10人小组中，所有人生日都不同的概率等于0.883（比2人或3人小组的概率要低得多），所以至少2人生日相同的概率是0.117（1–0.883）。

第一个分数是第二人和第一个人生日不同的概率。第二个分数是第三人和前两个人生日不同的概率。第三个分数是第四个人和前3个人生日不同的概率，以此类推。第九个也是最后一个分数是第十个人和任何其他9个人生日不同的概率。



所有人生日都不同意味着，一连串事件中的每一个事件都必须共同出现，所以我们通过将所有单个概率相乘来计算同一组所有事件发生的概率。每当我们添加一个人，我们就有一个分数进入方程，这使得最终乘积越来越小。

4.11.5 任意规模小组的解决方案

随着小组规模的增加，至少2人生日相同这一事件变得越来越有可能。这是非常合情合理的，但随着小组规模变大，至少2人生日相同的概率迅速变大的程度令大多数人震惊。图4-3说明当你添加越来越多的人时，概率上升的速率。

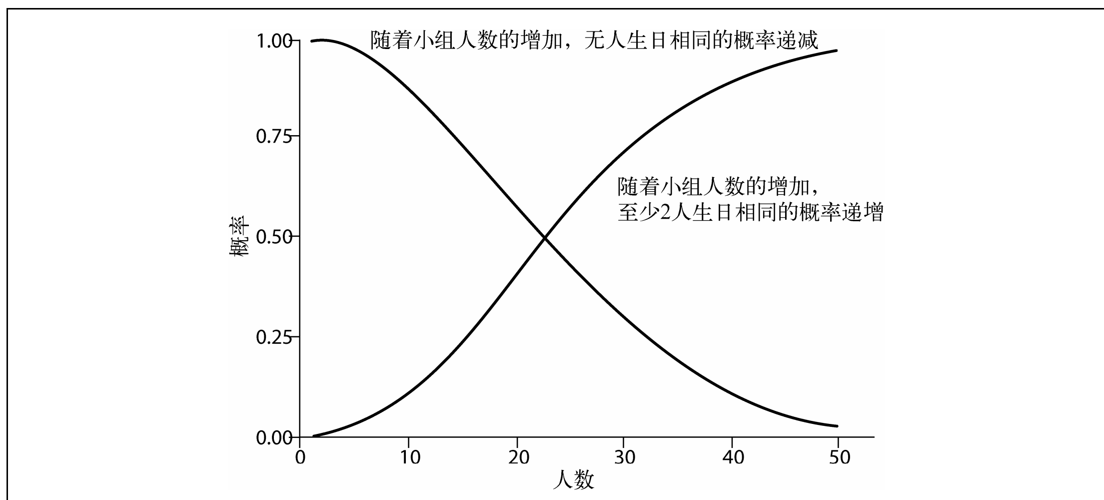


图4-3：生日相同的概率

对于20人来说，概率为0.411；30人的概率是0.706（即10次中有7次，你会在你的赌注上赢钱，这是相当不错的几率）。如果你组里有23人，至少2人生日相同的概率（0.507）只是比0.5稍微高一点。

不管怎么说，这是一个非常巧妙的方法，人们从未停止过对它的惊叹。但要记住，只有当房间里至少有23人时（并且你愿意接受50/50的赔率），你才可以下注。人越多，它越有效，因为每添加一个人，你获胜的几率就会显著上升。为了有90%的几率让你赢得赌注，你需要保证房间里有41人（至少2人同一天生日的概率等于0.903）。如果房间里有50人，你会有97%的几率赢得钱。一旦人数超过60，实际上你能保证房间里至少有2人生日相同，当然，如果有366人出席，至少2人生日相同的几率是100%。如果你能让别人和你打赌，这些都是很好的选择！

——威廉·斯科朗普斯基



4.12 设计你自己的酒吧赌局

通过一些计算，或许利用一些电子表格软件，你可以计算出各种各样“自发”友好赌注的概率。

本章其他几个统计Hack使用了纸牌[Hack #42]或骰子[Hack #43]作为道具，用来论证一些看似罕见和不寻常的结果实际上却是相当普遍的。作为对教育领域的统计原则感兴趣的一分子，毫无

疑问你希望使用这些教学实例来打动和指导他人。当然，如果你碰巧在此过程中赢了一点钱，也可视作教师生涯的好处之一。

但你没有必要完全依赖这里提供的具体例子，甚至随身携带纸牌和骰子（虽然，推己及人，你可能有很多其他的原因随身带着纸牌和骰子）。这里有几个基本的原则，你可以用这些原则对任何已知的数据分布制作自己的赌局，诸如字母、1~100的数字，等等。

- 原则1

一个不太可能的事件，如果有它有重复出现的机会，那么它出现的可能性会增加。

- 原则2

如果有大量的可能事件，那么发生任何特定事件的几率都很小。

接下来，本Hack会告诉你如何在自己的酒吧赌局里将这些原则转变为自身优势。

4.12.1　原则 1

任何给定事件发生的概率取决于匹配结果数，等于匹配结果数除以可能的结果数。例如，你和我在同一月份出生的概率是多少？迅速反应：出生月份平均分布在所有月份，概率为1/12。只有一个结果算作是匹配（你的出生月份），一共有12个可能的结果（一年有12个月）。

任何2个读这本书的人中，有人和我在同一月份出生的概率是多少？凭直觉应该大于1/12。很遗憾，这个概率的计算公式并不简单。比如，它本身不是1/12。这将产生一个比我们开始时（即1/24）更小的概率。也不是公式1/12+1/12，虽然2/12似乎有希望是正确答案，因为它大于1/12，意味着比之前有更大的可能性，但这种概率并不都是加法。为了证明简单地把两个分数相加是无效的，我们假设这个问题里有12个人。在12个人里找到一个和我出生月份匹配的几率显然不是12/12，因为这意味着肯定会有一个匹配。

一个事件在多次机会中出现的可能性的计算公式基于这样一个概念：每增加一次尝试（如骰子投掷），就多乘一次此事件不会发生的概率。这一过程结束后，用1减去这个结果，会得到该事件发生的概率。

这个公式有理论上的吸引力，因为它逻辑上等同于更直观的方法（它使用相同的信息）。它也具有数学上的吸引力，因为最终结果大于单一事件发生的概率，与我们的直觉相符。这样思考：有多少次它不会发生，而这些次中，有多少在下一次还不会发生？

下面是计算2个读者中，有人和我同一月份出生的概率公式：

$$1-\left(\frac{11}{12}\times\frac{11}{12}\right)=1-\left(0.917\times0.917\right)=1-0.841=0.159$$

4.12.2　原则 2

为了让别人接受你的赌注或用任何特定的结果让观众惊讶，从直觉上这个事件的可能性必须要小。所以，赌注或魔术可以与一年365天或一副扑克的52张牌有关，一本电话簿里所有可能的电话号码更有效、更惊人，因为和获胜结果数量（比如1）相比，这些数字看起来如此之大。

任何小概率事件在单一试验中发生的几率确实很小，所以这一原则中表达的直觉是正确的。但是，正如我们所看到的，如果进行多于一次的试验，该事件发生的几率会增加，并且可以迅速增加。

4.12.3　启动你的酒吧赌注

让我们遍历我刚制作的几个赌局，来证实我的优势。

1. 字母表中的字母

在这个赌中，我会从字母表里选5个字母。我敢打赌，如果我选择6人，并要求他们随机挑选任何一个字母，他们挑选的字母中有一个或多个会和我的5个字母里的字母相匹配。以下是投注展开方式。

- 可能的选择数

字母表中有26个字母。

- 单次尝试失败的概率

26个可能中有21个是不匹配的：21/26=0.808。

- 尝试次数

6次

- 6次尝试均失败的概率

$0.808^6=0.278$

- 6次尝试不均失败的概率

1−0.278=0.722

我赢得这个赌注的概率是72%。

2. 选择一个任意号码

这一次，我从数字1~100中选出10个数字。我敢打赌，如果我选择10人，并让他们随意从数字1~100挑选一个，他们挑选的数字中有一个或多个会和我的10个数字里的数字相匹配。以下是

计算过程。

- 可能的选择数量

有100个号码可供选择。

- 单次尝试失败的概率

100个可能有90个是不匹配的：90/100=0.90。

- 尝试次数

10次

- 10次尝试均失败的概率

$0.9^{10}=0.349$

- 10次尝试不均失败的概率

$1-0.349=0.651$

我赢得这个赌注的概率是65%。

3. 亲自实践

重复我刚刚展示给你的步骤和计算，开发自己的原创聚会技巧。所有这些都不需要任何道具，只需要一个有意愿且诚实的志愿者。

请注意，该计算基于人们随机挑选号码这一情况。当然，实际上，人们往往不会挑一个他们刚刚听到别人挑过的字母或数字。换句话说，他们的选择不独立于其他人的选择。如果当前的选择是基于之前的不正确选择（或不应选）而作出的，这将有助于提高你的胜算。例如，在100选10的数字赌中，如果10人中有人会选择别人选过的数字的可能性为0，那你获胜的几率将从65%上升到67%。

4.12.4 确保被骗的不是你

和别人玩是有趣的，但你永远不知道什么时候会落入别人设计的聪明统计陷阱里。例如，还记得你和我有相同的出生月份，从12个月里选1个月这样的几率吗？我骗了你！我出生在2月。那个月比其他月份的天数要少，所以你出生在2月的几率实际上小于1/12。2月有28.25天（偶尔出现的2月29日计作0.25），并且当年是365.25天（偶尔出现的闰年同样计数）。你和我出生在同一月份的几率是28.25/365.25（7.73%），而不是8.33%（1/12）。

所以，你不太可能和我有相同的出生月份。我想起来了，我的出生记录、出生证明等资料在很多年前的一场大火中不见了。所以，现在我出生的原始数据遗失了。

说不定，我可能甚至没出生呢！

HACK #47

4.13　疯狂地玩百搭牌

在扑克游戏中加入百搭牌，可以提高玩牌的乐趣。但是，从统计上来说，百搭牌也使得事情变得混乱了。

几百年前，扑克玩家制定了手牌的排名顺序，并决定什么牌可以战胜什么牌。令人高兴的是，从统计角度来说，他们制定的顺序和玩家发到的手牌概率是一个完美的匹配。据推测，扑克规则的开发者要么做了计算，要么引用了他们自己在实际玩牌中看到的每种手牌出现的频率。也可能是他们拿一副牌、纸和铅笔，在一个轻松的下午，自己随机发了成千上万次的扑克手牌，然后收集数据。不管用了什么方法，手牌大小的排序和特定组合牌的相对稀缺性是一个完美的匹配。

不过，排名顺序并没有考虑到某种类型的手牌和排名紧随其后的手牌类型之间意义重大的概率差距。例如，同花顺不出现的概率是排紧其后的四条的16倍，同时，顺子紧排在同花之后，而同花的出现概率只有顺子的一半。

在我们谈论玩百搭牌（指可替换持牌者任意想要值的牌，通常是王牌）的问题之前，让我们回顾一下扑克手牌的大小排序。表4-17所示的是在任意随机的5张牌里，出现一副给定手牌的概率；与排序紧随其后的手牌相比，每个手牌的相对稀少性也显示在下表中。

表4-17：扑克手牌、概率及对比

手牌	概率	相对稀少性
同花顺	0.000 015	16倍的不太可能
四条	0.000 24	5.8倍的不太可能
满堂红	0.001 4	1.4倍的不太可能
同花	0.001 9	2.1倍的不太可能
顺子	0.003 9	4.4倍的不太可能
三条	0.021	2.3倍的不太可能
两对	0.048	8.8倍的不太可能
一对	0.42	1.2倍的不太可能
什么都不是	0.50	----

对于赌徒来说，表4-17还有若干值得注意的观察结果。首先，5张牌的赌博中，有一半的时间玩家什么好牌都没有。几乎一半的时间，玩家有一个对子。玩家只有8%的时间获得比一对更好的手牌。

其次，一些出现可能性看似完全不同的手牌，出现的概率却几乎相同。注意，同花和满堂红出现的概率大致相同。

最后，三条之后，出现更好手牌的可能性迅速下降。事实上，在概率上出现了两个巨大的下

跌：大多数时候（92%）什么都没有或有一个对子，然后两对或三条发生的概率是另外的7%，只有不到1%的时间能看到比三条好的手牌。

4.13.1　百搭牌的问题

上面的结果非常有趣，但这到底和百搭牌的使用有什么关系？好吧，在一副牌里加入百搭牌把所有这些经过时间考验的概率都搞砸了。假设一名百搭牌持有人希望凑成最好的手牌，并且还假定已被把百搭牌添加到一副牌里了，表4-18显示了相比于传统概率的新概率。

表4-18：一副牌里有一张百搭牌时下列手牌出现的概率

手牌	百搭牌存在时的概率	经典概率	百搭牌导致的概率改变
五条	0.000 004 5	------	-------
同花顺	0.000 064	0.000 015	增加327%
四条	0.001 1	0.000 24	增加358%
满堂红	0.002 3	0.001 4	增加64%
同花	0.002 7	0.001 9	增加42%
顺子	0.007 2	0.003 9	增加85%
三条	0.048	0.021	增加129%
两对	0.043	0.048	减少10%
一对	0.44	0.42	增加5%
什么都不是	0.45	0.50	减少10%

通过新概率，我们可以发现百搭牌的问题很明显，尤其是当我们看三条和两对时。有了百搭牌后，三条比两对更常见！

传统上定好的手牌大小的排名顺序不再与实际概率相符。此外，当增加了一张百搭牌后，获得两对的几率实际上下降了。当然，其他概率也发生了变化，其他所有可玩的手牌变得更有可能。一些超级手牌，虽然依然罕见，但概率被极大地提高了：三条好的手牌出现的概率是之前出现概率的两倍左右。

知道这些新概率给精明的扑克玩家带来优势。经验丰富的专业扑克玩家认为百搭牌是幼稚的，是给业余玩家玩的，所以他们避免游戏时有百搭牌。事实上，和这种刻板印象相反，一些知情玩家会物色这些游戏，因为他们相信自己比你这种幼稚类型的人更有优势。（你懂的，那些幼稚类型的人，像是不读Hacks系列丛书的那些人？）

4.13.2　生效原理

正如你在表4-18看到的，使用百搭牌降低了获得两对的概率。但为什么会这样呢？无疑，增

加一张百搭牌意味着有时我可以把一对手牌变成两对手牌。这是真的，但为什么可以呢？想象一下，一个玩家手里有一对，他的第五张牌拿到了一张百搭牌。是的，他可以将那张百搭牌和一张单牌配成一对，声明有两对。另一方面，将这张百搭牌和他已有的一对进行匹配能得到三条，这是更聪明的做法。在考虑选择两对还是选择三条时，大家都会选更厉害的三条。

4.13.3　百搭牌的其他问题

百搭牌的存在，创造了让博弈理论家疯狂的一个悖论。悖论的原理如下。

(1) 手牌的排名和它们在扑克游戏中的相对价值应该基于其出现的频率。较少出现的手牌应该比较常见的手牌具有更多的价值。

(2) 在选择是否使用百搭牌把手牌变成两对或三条的情况下，玩家通常会选择组成三条。这实际上改变了频率，使得两对变得比三条更少见。

(3) 因为排名应根据概率制定，所以当有百搭牌参与时，应该将扑克规则变成两对，这比三条更有价值。

(4) 采用修订后的排名，三条的价值不如两对的价值，所以现在聪明的玩家会用他们的百搭牌凑成两对以代替三条，所以很快两对就会变得比三条更加常见。

(5) 排名规则将不得不再次改变以匹配因为之前规则变动所导致的实际频率，永无止境的循环将就此开始。

表4-18假设玩家会根据传统排名来组成他们最好的手牌，这样就避免了这一悖论。我很聪明，是吧？想玩牌吗？



4.14　永远不要相信一枚诚实的硬币

在通常世俗统计的世界里，所有的神圣事物中，没有什么比旋转一枚诚实的硬币更可信了。无论是正面还是反面，几率都是50%，对不对？显然令人不安的答案是：不是50%！

对几率以及运作原理的基本解释，几乎总是包含一个简单的抛硬币或旋转硬币的例子。“正面你赢；反面我赢。”是解决各种纠纷的通常方法，二项分布[Hack #66]通常作为随机硬币结果的样式来描述和讲授。

但事实证明，如果你旋转硬币，尤其是一枚全新的硬币，反面朝上的次数可能比正面朝上的次数多。

4.14.1　闪耀的新便士

你知道一枚全新的一分钱的外观和感觉，对吧？它是如此明亮，以至于看起来很假。它的细

节是如此丰富，它的边缘是如此锋利，因此你要小心，不要割伤自己。

好了，给自己准备一枚明亮、尖锐的小硬币，将它旋转100次左右。收集正面和反面的结果数据，准备震惊吧，因为出现反面的次数很可能会超过50次。如果我们对硬币公平性的理解是正确的，那么一枚硬币出现反面的几率是一半。（大声说最后一句，这样它更有意义。）但是，新便士不是这样的。

新硬币，至少新便士，往往有一个清晰的边缘，实际上在反面有点长或有点高（便士的反面比正面刻的要深一点）。图4-4给出了边缘看起来如何的概念。如果旋转这样一个形状物体，朝上的面往往会是长边。

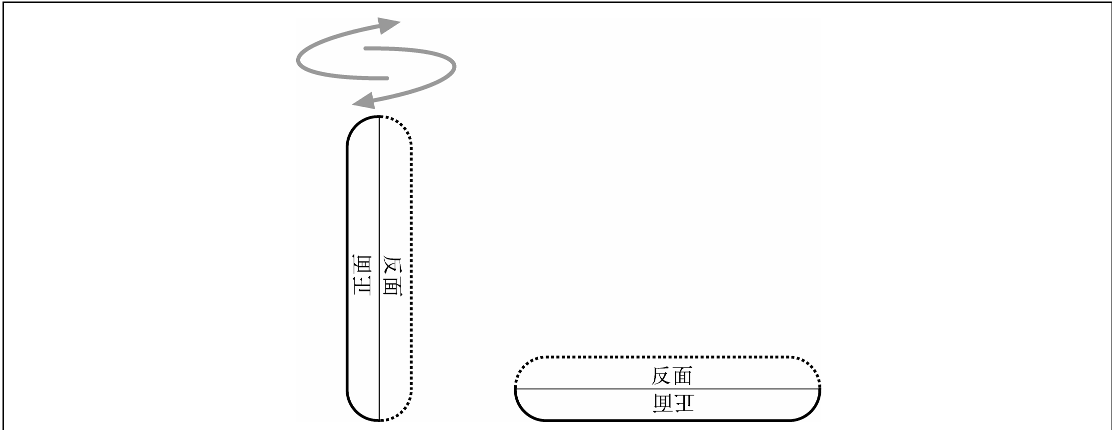


图4-4：旋转新便士

想象一下，旋转一个从啤酒或苏打汽水上取下的瓶盖。它不但不会转得这么好，而且你也不会因为看到它锋利面朝上而感到惊讶。新便士的形状有点像一个瓶盖，只是没有那么不对称。不过，就凭这点小小的锋利，如果旋转很多次，反面的优势就会体现出来。

4.14.2　二项式期望

可能存在的*瓶盖效应*提出了一个可检验的假说：

旋转一个刚铸造的便士，反面朝上的概率大于50%。

当然，几次旋转后，我们可能发现一枚硬币反面朝上的次数多于正面，或许这只是偶然，不能真的证明什么。我们知道，在小样本中出现某结果的概率不代表样本抽取总体的属性。

我们旋转硬币的例子应该代表无限次的硬币旋转。如果我们旋转硬币100次，发现51次是反面朝上，对我们的假设来说，这是可以接受的证据吗？也许不是；也无法解释0.50的比例。52次反面呢？一百万次旋转有52%的反面呢？

统计学再次来援救，并为我们的实验成果提供一个判断标准。我们从二项式分布知道，理论上一枚公平的硬币（没有不平衡的古怪的边）在100次旋转中，42%的时间产生51次或更多的反面。守旧派统计程序要求一个结果必须有5%或更低的出现几率，才被视为统计显著——不太可能是偶然发生的。因此，我们可能不会把100次旋转后，51%的反面出现作为对我们假设的支持。

另一方面，如果我们不断地旋转硬币6774次，得到了51%的反面，那随机发生的概率只有5%。这一结果的显著性水平为0.05。表4-19显示当预期的结果是反面的可能性为50%时，偶然获得某个比例反面的可能性。如果与这个预计比例的偏差有统计学意义，那么可以将它们视为支持我们假说的证据。

表4-19：硬币旋转和某个结果的概率

旋转次数	反面比例	给定比例或更高的概率
100	0.51	0.42
100	0.55	0.16
100	0.58	0.05
500	0.51	0.33
500	0.55	0.01
500	0.58	0.000 2
1000	0.51	0.26
1000	0.55	0.001
1000	0.58	0.000 000 2

请注意，这种分析能力真的会随着样本容量变大而增加[Hack #8]。如果你旋转硬币500次或1000次，只需要预期的小幅波动，就可以支持你的假设。而对于100次旋转，你需要看到反面的比例等于或高于0.58，这样才能证明新铸的便士确实有反面优势。

观测比例和预期比例的距离表示为z分数[Hack #26]。下面是计算z分数和生成表4-19中数据的公式：

$$z = \frac{\text{观测比例} - \text{期望比例}}{\sqrt{\dfrac{\text{期望比例}(1-\text{期望比例})}{\text{样本大小}}}}$$

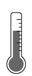
分配的概率是正态曲线下的区域，仍然高于z分数。

4.14.3　无效领域

一旦你向自己证明，反面优势是真实的，那在你跑去赢得各种疯狂的赌注之前听一下我的提醒。你必须旋转硬币！不要弹它。跟我说：旋转，请不要用手指弹。

4.14.4 参阅

一个有趣的网站提出“瓶盖效应”（bottle-cap effect）这一术语，并包含了对硬币反面高出的边缘的热烈讨论。该网站由加里·莱姆希（Gary Ramseyer）博士维护：http://www.ilstu.edu/~gcramsey/。



HACK #49

4.15 知道你的极限

人并不总是能作出理性的决定。当预期回报巨大、赔率也公平的时候，即使是聪明的赌徒，有时也会拒绝下注。圣彼得堡悖论（St. Petersburg Paradox）给出了一个相当公平的赌博游戏示例，完全正常的统计学家很可能不玩这个游戏，只是因为他们是人。

对于精明的统计赌徒，标准的决策过程涉及以下步骤：计算一个假设赌注的平均回报和成本，然后确定是否可能收支平衡，能赚到很多钱则更好。虽然一个人能生成几十个关于是否应该玩游戏的统计分析，但人类的心理感觉有时会占据主导，人们会拒绝接受赌注，只是因为感觉不对。

4.15.1 圣彼得堡游戏

圣彼得堡游戏大概有300年的历史。1738年，丹尼尔·伯努利描述了游戏的参数。下面是一些规则。

(1) 你提前支付一定的费用给我。

(2) 抛硬币。如果正面朝上，你赢了，我会付给你2美元。

(3) 如果不是正面朝上，我们会再次抛硬币。如果这次正面出现，我会付给你2^2（4美元）。

(4) 假如正面依然没有出现，我们再次抛硬币。第三次抛硬币正面出现了，那我付你2^3（8美元）。

到目前为止，这听起来很不错，对你来说更为公平。但它会变得更好。我们不断抛硬币，直到正面出现。当正面最终出现时，我付给你2^n美元，其中n是出现正面需抛硬币的次数。

至少从你的角度来看，这是个伟大的游戏。但这里有个要命的问题：你会为这个游戏支付多少钱？



圣彼得堡游戏以前可能并没有作为受欢迎的赌博游戏在俄罗斯的大街小巷流行，但当赌钱时，它就一直被用作思维如何处理概率的假设示例。它为早期统计学家分析“预期结果”在我们头脑中的运转原理，找到了理由。顺便说一下，关于这一内容的论文实际上是由圣彼得堡科学院发表的，因而得名。

决定你会出多少来玩游戏是个有趣的过程。作为聪明的统计学家，你当然会支付不到2美元

的费用。即便没有得到更大回报的可能性，赌你会在第一次抛硬币时得到正面，也可以得到多于游戏投入的回报，这显然是一个不错的赌注，尝试一下吧。

您也可能会乐意支付2美元。你将有一半的时间赢回这2美元，而另一半的时间你会得到比这更多！这是一个可以保证你最终获胜的游戏，所以获胜不是问题。当你第一次没有得到正面时，你已经保证自己至少会赢4美元，甚至更多。

所以，也许你会支付4美元来玩这个游戏。当然，你的回报偶尔会非常大——8美元、16美元、32美元、64美元……理论上，回报可能接近无穷大。但是你会支付多少？这就是64美元的问题。

4.15.2 统计分析

一些社会科学研究人员认为，大多数人会花4美元来玩这个游戏，可能还会多一点。很少有人会出太多钱玩这个游戏。但是，从统计学角度分析，结果会是怎样的呢？你最多应该出多少钱？

好吧，我考虑上交我的统计粉丝俱乐部会员卡，因为我怕告诉你正确答案。由于涉及赌博，概率的规则建议人们应该不惜一切代价玩这个游戏。是的，一个统计学家会告诉你应不惜一切去玩这个游戏！只要成本没有达到无穷大，从理论上说，这就是一个好的赌注。

让我们算算。下表是前6次硬币翻转的回报：

翻转	可能性	游戏比例	赢得	预期支付
1	1：2	0.50	2美元	1美元
2	1：4	0.25	4美元	1美元
3	1：8	0.125	8美元	1美元
4	1：16	0.062 5	16美元	1美元
5	1：32	0.031 25	32美元	1美元
6	1：64	0.015 625	64美元	1美元



预期收益是在所有可能的结果中，你会赢得的平均金额。对于单次抛掷，有两种结果：正面，你就赢了2美元；另一种可能性，反面，你就会得到0美元。平均支出为1美元，一次硬币抛掷（事实证明，对于任何次数的硬币抛掷）的预期收益是1美元。

如果你玩这个游戏64次，你只在第六次掷硬币中获得正面，但你将赢得64美元。64次中的32次，你只会赢得2美元。平均收益听起来较低：才1美元。但是偶尔会出现这种情况：很长一段时间内都没有出现正面，当正面终于出现时，你已经赢了很多钱。当你开始游戏时，你不知道它会

持续多久，你也不知道它可能会持续很长时间（像彼得·杰克逊[7]的电影那么长）。

关于这一系列的投掷，以及几率随着奖金上升而以同样速率下降，有一些事情需要注意。

- 本表只显示了掷6次硬币的情况。不过，从理论上讲，投掷可以永远进行下去，一直不出现正面。
- 每掷一次硬币，奖金数额增加一倍，游戏中的投掷数量减半。
- “游戏比例”列的数值永远不会增加到1.0或100%，因为总是有一些偶然的机会，不管多么小，仍需要再掷一次。

我们的统计粉丝俱乐部会员决定是否玩赌博游戏的决策规则是：游戏的预期值是否大于玩的成本。预期值是通过把所有可能结果的预期回报相加计算出来的。

你应该记得每一个可能试验的预期收益为1美元。有无限数量的可能结果，因为硬币可能永远地不停地投掷，一直不出现正面。为了得到预期的价值，我们把这一系列的无穷的1美元相加，得到一个巨大的总和。对于这个游戏的期望值是无穷大的。因为当玩游戏的成本低于预期值时，你就应该玩这个游戏。只要玩这个游戏的成本还没达到无穷大，你就应该玩。

4.15.3 无效原因

当然，在现实生活中，人们不会为这样的游戏支付远超2美元的钱，即使他们知道所有的统计数据。没有人明确地知道为什么聪明的人为这样有前景的游戏付很多钱感到厌恶，但这里有一些理论能解释这一现象。

1. “无限”是很多

即使你在精神上接受，从长远来看比赛是公平的，玩很多、很多次的话偶尔也会得到很大的回报，但是“长远来看”是无限长的，这是一个相当长的时间。很少人有耐心或有足够多的钱来玩这样一个需要这么多耐心和费用的游戏。

2. 边际效用递减

这个问题的鼻祖伯努利认为，人们将金钱视为是有价值的，但这种观念不和金额成正比。换句话说，虽然16美元优于8美元，但16美元和8美元的相对价值，和128美元与64美元的相对价值，是不一样的。

因此，在某些时候，作为奖励的金钱无限翻倍不再有同样的意义。伯努利还相信，如果你有很多钱，和你有很少的钱相比，一个小赌注的意义对后者来说更大。（有点像那些腰缠万贯的卡通人物用百元大钞来点雪茄。）

注7：《魔戒》的导演。——译者注

3. 风险与报酬

人类往往倾向于风险规避。也就是说，它们会偶尔冒险以换取报酬，但他们希望这种风险和成功的几率相符。圣彼得堡游戏有获得巨奖的机会，这是事实，但和风险相比，这个机会可能被视作太小，即使是4美元的风险。

4. 无穷不存在

有些哲学家会说，人们不把无穷的概念视作具体存在。任何通过鼓吹回报无穷大以鼓励人们玩游戏的摊位，都不怎么引人注目。

这也许就是我不买彩票的原因。我不玩彩票，因为通过买彩票，我获胜的概率只增加了一点点。对于我来说，我中奖的概率是无限小，或非常接近无限小，以至于我不把获奖的可能视作现实。

4.15.4　参阅

- “明智地下注”[Hack #35]。
- 《斯坦福哲学百科全书》（*Stanford Encyclopedia of Philosophy*）中有关于圣彼得堡悖论的有趣且思虑周全的讨论。网址为：http://plato.stanford.edu/entries/paradox-stpetersburg。

第 5 章

游戏技巧

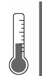
（*Hack #50~#60*）

并非只有赌博游戏才有统计数据。你可以使用游戏专用概率知识，在电视真人秀[Hack #50]、大富翁游戏[Hack #51]或指导足球队[Hack #58]中取得胜利。

你在日常生活中，最常见到统计的地方可能是体育领域，虽然“统计”一词和统计Hacker使用它的方法并不完全一样。体育迷们往往把*数据*视作*统计*。无论如何，有大量的Hack可以帮助你在比赛结束前预测比赛结果[Hack #56]，甚至在比赛开始时预测结果[Hack #55]。

历史是对未来的最佳指南，最好的预测需要用不同的方法来对球队和球员的以往表现进行跟踪、可视化[Hack #57]和排名[Hack #59]。

当然，如果你是一名真正的统计黑客，那么你可以想出一些统计游戏，比如用椰子建立一个会学习的电脑[Hack #52]，通过邮件玩纸牌把戏[Hack #53]，让你的iPod保持诚实[Hack #54]，或随机估计圆周率的值[Hack #60]：它们都非常有趣。



HACK #50

5.1 避免筋疲力尽

在电视真人秀节目Let's Make a Deal中，参赛者总是在3面窗帘之间进行选择。对于这些类型的情况，有种统计策略能帮你赢得别克汽车，而不是永远都吃不完的Rice-A-Roni[1]。

试想一下，如果你愿意，你和叔叔弗兰克一起旅行时正经过堪萨斯通加诺克西（Tonganoxie，Kansas）的未知区域。你们走到了一个岔路口，这个岔路口分出3条可行的路：A、B和C。你们不知道哪条岔路口通往目的地：传说中世界上最大的麻线球（在堪萨斯州的考克市）。一位年迈

注1：一种盒装食物，里面包含大米、意式细面和调料。——译者注

的探矿者和他的毛驴在十字路口休息。

“喂，老人家，”你说，“这条路通往世界上最大的麻线球吗？”

“嗯，”他说，“我知道，但我是不会告诉你的。不过，我可以告诉你其中有一条是正确的路。剩下两条是错误的，通向某些灾难（或者至少是年久失修的厕所）。前进吧，随你挑，时髦的城里人。你往前开，回头看我，我不会给你走对了还是走错了的暗示，但我会指向另外两条道路的其中一条。我指出的那条路是错误的。当然，你仍然不会知道你是否猜对了，但我保证，我指出另外两条路中的那条是错误的。”

你接受了这个陌生男人的建议（你有的选吗），让弗兰克叔叔这个比你经验丰富的赌徒挑选道路。他随机选了一条，你乐观地走向了3条路中的一条——假设是A。当你回头看时，好心的探矿者指向了其他两条道路的一条——假设是B。你马上踩刹车，车子猛然停了一下。你不顾弗兰克叔叔的反对，朝剩下的道路C前进，并坚信现在走的是正确的道路。

疯了，是吗？发烧把脑子烧晕了？不，你刚刚应用了统计方法来解决知名的蒙提霍尔问题（Monty Hall problem），并从3条路中选择了最有可能正确的一条路。难以置信，对吗？继续往下读，我的朋友，准备赢得比你最疯狂的梦想还大的财富吧。

在这种情况下，最好的策略是违背直觉且非常怪异的，以至于世界上最聪明的人也不认为它真的是好的甚至是最好的策略。但相信我，它是。

5.1.1　蒙提霍尔问题和真人秀策略

在我们的3条道路和探矿者的例子中，事实上，C是正确道路的几率是2/3（67%）。若要将此策略应用到更真实的情况中，想想真人秀节目中的选手或是任何游戏中的赌徒，这些游戏的奖品隐藏在盒子或门后。由于真人秀理论家和思维活跃的统计人员对其进行了深刻的探讨，这个问题在真人秀节目*Let's Make a Deal*中相当普遍（20世纪60年代至70年代是它的全盛时期），但它仍然能在如今的电视真人秀节目中看到。*Let's Make a Deal*的主持人是蒙提霍尔，这个问题以他的名字命名。

在真人秀情境下，这个问题是这样的。蒙提给你呈现三面窗帘。他知道每一个帘子后面是什么。他解释说，一面窗帘背后是辆新车，其他两面窗帘后面是不值钱的奖品，蒙提把它们称作zonk。（zonk往往指驴或巨大摇椅这类东西，没有任何真正用处。）他让你挑选一面窗帘，不管它后面是什么，你都将赢走它后面的东西。比方说，你挑了窗帘A。然后，他打开一个你未选择的窗帘，比如B，它后面有一个zonk。然后，他提供给你一个机会，你可以放弃原来的选择，转而去选剩下的窗帘C。你应该改变选择吗？

和3条道路那个问题一样，答案是肯定的，你应该改变选择。第一次听到这个答案时似乎感觉不太正确。但是，如果你想要提高赢得汽车的概率，你就应该改变选择。

5.1.2 为什么应该改变选择

想想你猜中窗帘的概率。我们假设它是一个随机的猜测——没有其他因素的干扰，比如，我注意到一面帘子动了，我认为它背后有一头驴在跳。

3面窗帘，只有一面帘子是正确答案，这意味着你有1/3的几率猜中，从而赢得汽车。这大约是33%。在第一次猜测时，没有额外的信息，你可能会错；事实上，你有2/3的几率会错。换句话说，有大约67%的几率，这辆车在你没有挑选的那两面窗帘的后面。

你知道另外两面窗帘中，其中一面的后方一定没有汽车，但这不会改变这辆车可能位于某面未被选择的窗帘后面的概率（67%）。记住，不管你选择哪一个，蒙提永远都会打开一面错误的窗帘。这辆车在B或C窗帘后面的概率为67%，这仍然正确，即使B被揭开后发现它后面没有汽车。67%的概率现在变成窗帘C了。这就是你为什么应该改变窗帘选择的原因。



如果给你机会，你可以把已选的窗帘换为另外两个未选的窗帘，你会立刻换，不是吗？这就是蒙提霍尔难题的本质所在。

为了打消你内心深处的怀疑，我们可能还需要一些数字支撑。看一下表5-1，它展示了游戏最开始3个选项的概率分解。你有1/3的几率猜中、2/3的几率选到不能获得大奖的窗帘。

表5-1：游戏开始时汽车所在位置的概率

窗帘A	窗帘B	窗帘C
33.33%	33.33%	33.33%

表5-2以不同的方式显示了相同的概率分布，但它并没有改变问题的任何参数。

表5-2：另一种关于游戏开始时汽车所在位置的概率表述

窗帘A	窗帘B或窗帘C
33.33%	66.66%

表5-3显示了蒙提揭示你未选择窗帘之一（窗帘B）不是正确窗帘后的概率。67%的可能性现在转移到窗帘C上了。

表5-3：窗帘B被揭开后汽车所在位置的概率

窗帘A	窗帘B	窗帘C
33.33%	0.00%	66.66%

在任何相似情况下，你都应该换窗帘。当然，你可能是错的，但如果你接受提供的交换机会，

那么你就有一个更高的几率来赢得汽车，或其他任何你在玩的游戏的奖品。如果满足下列标准，这永远是最好的策略：

- 主持人知道每面帘子后是什么；
- 主持人揭开你未选择的窗帘之一，奖品不在这面窗帘后面；
- 你当初的选择是随机的。

即使这个解决方案的正确性不能立刻体现出来，也不要过于担心。真正聪明的人往往会关注两面尚未揭开的窗帘，并将新的概率看作50/50，因此，不管你是否更改选择，都没有关系。但是，要记住关键的一点：你最初挑到正确窗帘的几率为33.3%，不管你作出决策后发生了什么，这个几率都不会改变。虽然专家有时不同意这是思考此问题的最佳办法。但即使是像本章开头提到的你在经过通加诺克西时遇到的探矿者一样聪明的人，也不总是知道蒙提霍尔问题的正确答案。他赢得了那头驴，你怎么看？

争　　议

蒙提霍尔问题以及由此导致的通用真人秀策略，最初由*Parade*杂志的专栏作家玛丽莲·沃斯莎凡特（Marilyn Vos Savant）于1991年介绍给大众。因为她被称为“高IQ天才”，沃斯莎凡特回答读者的问题，有时甚至是脑筋急转弯问题。有人把我刚描述的问题发给了她，她发表了我在这里给出的答案。

显然，她收到了许多信件，有些信表达了笔者的愤怒。这些愤怒的信来自于统计学家、哲学家和声称她错误的人。在学术期刊里，甚至出版了关于她的回答是否正确的争议。我对争议的看法是：事实证明，大部分的争论集中在问题的关键部分——蒙提知道每扇门后面是什么，所以当他打开第一面窗帘时，他知道后面是zonk。否则，揭开一面窗帘不算作新信息，沃斯莎凡特给出的答案也值得商榷。对她答案的大部分批评忽略了原来出版问题的一部分。



5.2　经过 GO 方格，取得 200 美元，赢得比赛

大富翁是一种几率游戏（几率卡）。因此，赢得游戏的最佳策略是充分利用概率。

要想赢得非常流行的帕克兄弟公司（Parker Brother）的棋盘游戏大富翁，需要谈判的技巧、巧妙的金钱管理和有见地的投资规划。还需要一点点运气。

由于两个六面骰子（以及随机洗牌的一堆卡牌）是决定你落在哪个方格的决定性因素，所以运气对结果的影响不只一点点。争强好胜的统计学家，比如你和我（至少是我）被所有概率在其中起着关键作用的游戏所吸引，其原因是，通过应用一些概率基础知识，我们会比平时一般情况下赢的次数多。

5.2.1　大富翁统计基础知识

让我们先分析掷两个骰子的简单效果。图5-1显示了每个人在第一回合中，最常落入的方格。

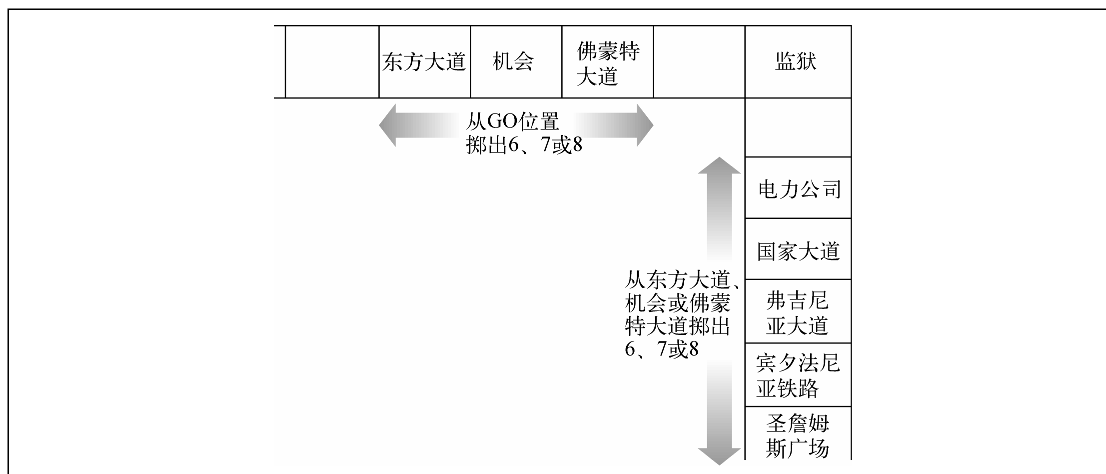


图5-1：开局可能落入的方格

想象一下，在游戏的开始，每个人都在GO的位置。两个六面骰子，有44.5%的几率，会扔出一个6、7或是8，其中7是最有可能的结果（16.7%）。那么，对你的第一次投掷，有一些方格更容易落入（例如，浅蓝色和弗吉尼亚大道），有一些方格不那么容易落入（例如，波罗的海大道或所得税）。仅根据开局的骰子投掷来看，不是所有的方格都有等同的落入机会。



当从GO开始时，甚至都无法落入地中海大道，因为投掷两个骰子得到1是不可能的。你有没有发现，地中海大道几乎总能成为在最后仍然可以购买的地产？

GO方格是一个很好的起点，用于计算落入方格的各种概率。不仅因为每个人都从那里开始，也因为那里会给玩家发一张机会卡。另一方面，如果一名玩家击中了“进监狱”（Go to Jail）这个方格，那么他就直接进监狱了，绕过了GO方格。所以，落入GO方格的概率不仅受骰子投掷可能出现的排列组合的影响，同样也受到各种机会卡的影响。机会卡会把玩家送到不同的地方。落入GO方格的概率还受到游戏本身规则的影响，其中包括能产生某些后果的方格、进监狱的情况以及出监狱的情况。

5.2.2　关键地产

我一直以GO方格为例，但是，当然，GO甚至都不是一个我们可以购买的方格。我们真正想知道的是购买什么地产，交易什么地产和先建哪块地。我们想要的是人流较多的区域；房地产成

功的秘诀是“位置，位置，位置”（显然，这些房子都有不错的木质露台或门廊，其中的原因我无法解释）。

表5-4显示了考虑所有规则的情况下，最常落入的前20个方格。该表还显示了一名玩家落入其中一个方格的几率。请记住，方格“平均”有2.5%几率是你最后落入的位置（40个方格除100是2.5）。

表5-4：大西洋城中最好的不动产

方　　格	排　　名	在此方格结束你回合的几率
监狱	1	11.60%
伊利诺伊州大道（Illinois Avenue）	2	2.99%
GO	3	2.91%
B & O铁路	4	2.89%
免费停车	5	2.83%
田纳西州大道（Tennessee Avenue）	6	2.82%
纽约大道（New York Avenue）	7	2.81%
雷丁铁路公司（Reading Railroad）	8	2.80%
圣詹姆斯广场（St. James Place）	9	2.68%
自来水厂（Water Works）	10	2.65%
宾夕法尼亚大道（Pennsylvania Avenue）	11	2.64%
肯塔基大道（Kentucky Avenue）	12	2.61%
电力公司（Electric Company）	13	2.61%
印第安纳大道（Indiana Avenue）	14	2.56%
圣查尔斯广场（St. Charles Place）	15	2.56%
大西洋大道（Atlantic Avenue）	16	2.54%
太平洋大道（Pacific Avenue）	17	2.52%
文特诺大道（Ventnor Avenue）	18	2.52%
浮桥（Boardwalk）	19	2.48%
北卡罗莱纳州大道（North Carolina Avenue）	20	2.47%

表5-4中的信息来自Truman Collins网站，网站地址是http://www.tkcs-collins.com/truman/monopoly/monopoly.shtml。聪明的柯林斯先生不但开发出了概率树，还用电脑模拟验证这些值，他为这些验证的值提供两种情况：玩家想尽可能长时间留在监狱（为了赚取租金，同时不必支付租金）；玩家希望尽快走出监狱（为了购买还可以买的地产）。我报告的值适用于前一种策略。

你可以从这些数据中得出一些重要的战术结论：

- 充分利用监狱

你的对手会有显著的12%的几率在监狱方格里开始他的回合。显然，持有以及开发刚刚被释放的玩家最有可能落入的土地，是一个明智的目标。这意味着橙色的地产（圣詹姆斯和他的兄弟），其次是红色（例如，伊利诺伊大道）和紫色（圣查尔斯和他的朋友）。

- 持有橙色

橙色地产共有3处，且都排在前10位。大约投掷骰子12次会有1次落入田纳西州、纽约大道或圣詹姆斯广场。垄断这些地产并快速发展，这似乎是名纯粹统计学家会选择的战略。

- 避免远边

棋盘远边的地产——绿色、浮桥和停车位置都不太可能落入，即使游戏进行了很长时间。只有浮桥和太平洋大道排名靠前，毫无疑问，浮桥之所以排名靠前是因为送玩家机会卡。这些地产开发起来最昂贵，所以将这些垄断纳入游戏计划有点冒险。

5.2.3 大富翁监狱系统的重要性

如果没有统计分析，监狱和“进监狱”方格在房地产整体价值中发挥的作用可能就不是那么清楚。玩家希望监狱可以销售。玩家在监狱方格里开始或结束自己回合的频率，比他们落入棋盘上任意垄断地产的频率要高。络绎不绝的刑满释放人员如洪水一般穿越棋盘的一边，这增加了去往伊利诺伊州一路上收取物业租金的几率。

玩家在经过街头时必须给其他玩家支付租金，这时候监狱也可以提供一个受欢迎的喘息机会，但是在游戏初期，监狱会阻止你购买你梦想的地产。监狱重要性的最后一点：只有一个方格，你永远无法结束你的回合。你能说出它来吗？进监狱（Go to Jail）。

5.2.4 参阅

- 比尔巴特勒运行另一个网站，介绍与大富翁相关的概率：http://www.durangobill.com/Monopoly.html。此外，该网站举办了一个讨论：当一个人希望他的计算涵盖大富翁游戏的每一个现实细节时，涉及的困难难以想象。比如，追踪一张特定的机会卡或公益卡是否发出。
- 落如一个方格的概率计算公式（在这个例子中用英国伦敦、英格兰、街道名称），可以在这里找到：http://hometown.aol.co.uk/monopolycheat/prob/method.html。



5.3 使用随机选择的人工智能

在微处理器发明之前，统计学家已经能够构建智能的、有学习能力的电脑。你可以使用椰子壳和概率规律来建造一台会学习的、在井字游戏上永不输的电脑。

有一个笑话说的是20世纪60年代的电视节目《吉利根岛》（*Gilligan's Island*）：教授总是用椰子和藤蔓制造电脑、洗衣机或火箭船。我感觉制造洗衣机和火箭船听起来不切实际，但是漂流者完全可以用椰子制出电脑。你也可以做到。如果你曾经被困在一个荒岛上，想要有个同伴，造一部电脑吧。

你不用像《荒岛余生》（*Castaway*）里汤姆·汉克斯扮演的角色一样想要一个排球，排球没有什么个性，但你的电脑能和你一起玩游戏，甚至会学习，并能因此变得更聪明。学习算法背后的驱动力来自几率和随机选择。

5.3.1　试误学习

据行为心理学家分析，所有动物（包括人类、水獭和单细胞生物）的学习方式都基本相同。经验告诉我们不同选择导致不同结果。由于动物接收结果的反馈，所以它就适应了。如果结果是积极的，这个生物更有可能在不久的将来作出同样的选择。如果结果是消极的，这个生物不太可能再一次作出这样的选择。

请注意，我们并没有保证一个“好”的行为总是会反复进行，或保证不良的行为会逐渐灭绝，它只是概率问题。动物更可能作出正确的决策，而不太可能作出错误的决策。为了使一台机器模仿动物的学习方式，我们必须从这个概率角度来建立机器。

玩游戏反映了很多试错的学习过程，因为结果很容易被解释为积极的（赢）或消极的（输）。在游戏中，反馈往往是即时的，而研究表明，选择和反馈之间的时间接近程度是学习（learning）是否发生的关键因素。请记住，学习在这里被定义为：正确选择的可能性增加或不正确选择的可能性减少。

5.3.2　建立一个井字机器

被困在岛上时，没有朋友的你或许希望通过和智能对手玩游戏来打发无聊时间。下面是建立一个不使用任何电或硅的奇妙装置的指令，这个装置会玩游戏，并具备像样的竞争力。

这款机器会学习：你和它对战的次数越多，它就变得越强。这个机器玩的是井字游戏，但理论上来说，你可以使用同样的原理来建立任意的双人战略游戏装置。井字游戏很简单，它很好地展示了设计、制造和操作方法。

如果《吉利根岛》里的教授曾经用椰子制造出一部电脑，他很可能受到生物学家唐纳德·米基（Donald Michie）的开创性工作和他那火柴盒的影响。1963年，米基在《电脑杂志》（*Computer Journal*）的第一期发表了一篇文章，吉利根和他的好朋友被困在岛屿上是若干年后的事了。米基介绍了他是如何设计的，并的确使用下列完整列表制造了一个不用电的电脑。

- 287个火柴盒

火柴盒有个可以打开的小抽屉。米基在每个火柴盒上标记出井字游戏中可能出现的287种不同结果中的一种。其实有更多的可能位置，但由于3行3列的标准井字布局是对称的，因此4种不同特征的位置可以只用一个位置来概括表示。在游戏中的每一时刻，“棋盘”的当前布局将操作人员引导到相应的火柴盒。

- 大量供应的9种不同颜色的珠子

这9种颜色代表井字棋盘的9个不同空间。最开始，每一个火柴盒里的珠子数等于下一步的移动数，且珠子颜色和可移动空间对应。只有代表合法移动的珠子会被放在相应盒子里。当然，不同的位置和火柴盒，只对应一小组合法的下一步移动，所以每个盒子里都混杂着不同颜色的珠子。

教授可能会用椰子壳代替火柴盒，用沙卵石或种子（或许用豪威尔先生存钱罐里的钱，他一直随身携带存钱罐）代替珠子。从你所处的热带环境里收集这些物资，将卵石填充的椰子进行有效分组，你就有了可以在荒岛上玩游戏的电脑。是的，你需要费力找到287个椰子，但你有其他更好的事情可做吗？

5.3.3　操作电脑

为了和你那“卵石供电”的电脑玩游戏，请按以下说明进行操作。

(1) 电脑先走。找到标记有当前位置的椰子。（对于第一步，这是一个空白的布局。）闭上你的眼睛并随机抽出一块卵石。

(2) 在你的棋盘上，在卵石颜色指示的地方标记一个X（我假定是在沙上画）。将卵石放在一个安全的地方。

(3) 作出你的移动，在你选择的地方标记一个O。

(4) 目前在棋盘上有一个新位置了。转至相应的椰子，从里面随机抽出一块卵石。回到第2步。

(5) 重复步骤2至步骤4，直到有一人胜出或平局。

接下来发生的事情是最重要的部分，因为它能让电脑学会如何玩得更好。行为心理学家把这种最后阶段称作强化（reinforcement）。

如果电脑输了，你通过把从椰子中随机抽取的卵石扔到海里来“惩罚”它。

如果机器赢了或打成平局，将鹅卵石放回到它们原来所在的椰子里，并通过额外加入一颗颜色相同的卵石来“奖励”它。

5.3.4　生效原理

奖励或惩罚电脑的过程基本上复制了动物的学习过程。积极结果导致奖励行为可能性的增加，而消极结果导致惩罚行为可能性的降低。通过添加或删除卵石，你的确是增加或减小机器在游戏中作出某种移动的真实可能性。

考虑游戏进行到如下阶段，电脑的移动用X表示，现在电脑必须走棋：

X	O	X
	O	

你可能意识到最好且唯一的可行走棋是电脑把X放在底部中间来阻止你获胜。但是，电脑意识到几种可能性。它考虑任何合法的移动。电脑考虑的两种走棋方法（这实际上意味着，它将允许被随机地从椰子壳中取出来）一个是最好的走棋，一个是最坏的走棋：

X	O	X
	O	
	X	

X	O	X
	O	
X		

如果电脑第一次玩这个游戏，这两种走棋（或行为）发生的可能性等同。在这种情况下，其他走棋也有可能，它们发生的可能性也相同。左边的走棋可能不会导致失败，至少不会立即失败，所以代表那步移动的卵石被添加到椰子中，相对于其他走棋，这种走棋的相对概率增加了。右边的走棋很可能以失败告终（除非和吉利根比赛，也许吧），所以这种走棋下次被选中的几率在数学上减少了，因为可供随机选择的这种颜色的卵石数量变少了。

任何给定的走棋被选中的概率可以通过这个简单的公式表示：

$$\frac{\text{表示走棋的卵石数量}}{\text{对应当前棋盘布局的椰子中的卵石总数}}$$

机器开始时有相等数目的卵石，或者，换句话说，任何一系列的走棋被选中的概率相等。当然，一些走棋在我们经验丰富的玩家眼里是非常愚蠢的，在真正的游戏中绝不会作出那些愚蠢的行为，除了非常幼稚的玩家。但是行为心理学家争论的问题是：所有生物，在它们建立一个大型的经验池前都是新手。这种经验池塑造了它们行为中的基础概率。

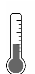
5.3.5　剖析本条Hack

可以用几种方法修改你的机器，使它变得更聪明。例如，对平局和获胜采用不同的奖励方法。这应该会更快地培养出一个好玩家。米基建议获胜奖励三个珠子，平局奖励一个珠子。

如果你想模拟动物的学习过程，那你可以调整系统，使临近游戏结束时的走棋比开始时的走

棋更加重要。这是为了反映这样一种观察：当强化最接近行为发生时，强化是最有效的。在井字游戏中，对于导致立即输掉的错误，应予以更有效的处理和惩罚。在游戏后期，随着用来走棋的珠子越来越少，学习的发生会更快。

一个明显能使电脑变得更聪明的升级是：甚至不容许电脑下坏棋，即不把代表会导致立马输掉的卵石放到你的容器里。这会解决电脑开始时智力低下的问题，但它并没有真正体现动物的学习方式。所以，虽然这可能是一个强大的竞争对手，但教授会因你缺乏科学的严谨性而感到失望。


HACK #53

5.4　信件传递的卡牌伎俩

按理说，洗好的纸牌应该是随机的。科学分析表明它实际上不是随机的，你可以充分利用卡牌分布的已知概率来对陌生人展现一个惊人的纸牌戏法。

想象一下，你在邮箱里收到一个厚厚的、神秘的信封。你没有将它交给最近的国家安全人员处理，而是打开了它，你在里面发现了一副普通的扑克牌以及下面一组说明：

(1) 切牌；
(2) 用交叠式洗牌法（在本Hack后面会定义）洗一次牌；
(3) 再次切牌；
(4) 再次使用交叠式洗牌法洗一次牌；
(5) 再次切牌；
(6) 取下卡牌顶部的那张牌，把它记下来，并将其随机放回卡牌里；
(7) 再次切牌；
(8) 重新洗牌；
(9) 再切牌一次；
(10) 把这副牌邮寄回附上的地址（在堪萨斯州的通加诺克西，或其他一些让人想起奇迹和奇思妙想的地方）。

你遵循这些说明（同时还戴着防护橡胶手套），把这副牌邮寄回去。大约一周过后，你收到一个小信封。里面正是你选择的那张卡！（也有可能是300美元的请求和预测你未来的邀约，此时的你只需扔掉那个邀约即可。）

令人惊异，是吗？不可能的，你说呢？混洗卡牌的已知可能分布使这变得很有可能，甚至像你这样的初出茅庐的统计学者也可以做到这一点，都不需要报名到霍格沃茨学校（Hogwarts）[2]学习。

注2：《哈利·波特》里的魔法学校。——译者注

5.4.1 生效原理

从数学上来讲，大家已经熟知一副扑克牌各种类型的洗牌效果。虽然彻底的洗牌（如燕尾或交叠式洗牌，使卡牌的两半交织在一起）是为了真正把一副牌洗成完全不同于原有次序的新次序，但是即使经过多次切牌和洗牌，原始卡牌的部分序列，依然保持着原有的秩序。

统计学家已经分析了这些模式并把它们发表在学术期刊上。这工作类似于这样的开创性建议，即为了在下一轮手牌前获得扑克、黑桃或桥牌的最佳组合，应该洗正好7次牌。

想象一副以某种次序排列的扑克牌。一轮洗牌后，如果洗牌是完美的，我们仍然可以在混合分布的卡牌里发现原来的次序。事实上，现在的次序是两种原来次序的相互重叠，并且通过交替选牌，你可以重构原来所有的次序。

表5-5显示了一副扑克牌进行一次完美洗牌的前后情况。为高效起见，只显示了12张牌，但这些原则适用于一副完整的52张扑克牌。

表5-5：完美洗牌对卡牌分布的影响

洗牌之前	洗牌之后
1. 方片A	1. 方片A
2. 方片2	7. 方片7
3. 方片3	2. 方片2
4. 方片4	8. 方片8
5. 方片5	3. 方片3
6. 方片6	9. 方片9
7. 方片7	4. 方片4
8. 方片8	10. 方片10
9. 方片9	5. 方片5
10. 方片10	11. 方片11
11. 方片11	6. 方片6
12. 方片12	12. 方片12

如果知道这12张牌的开始顺序，你可以在新牌组里每隔一张进行查看，就能够相当容易地把它挑出来。这些子模式的特点是保持上升的序列：当你沿着卡牌顺序移动时，卡牌的面值在上升。如果卡牌以一个很长的上升序列（或者4组，因为有4种花色）开始，鸽尾洗牌也将保持这些上升序列，它们只是交织在一起而已。即使经过多次洗牌，鸽尾洗牌也将保持这些上升序列。

如果在洗牌和切牌过程中的任意时刻从卡牌中抽取一张牌，并有目的地插入卡牌的其他地方，和总体的上升格局序列相比，这会出现“出位”（out of place）的情况。当然，这也正是卡牌伎俩的说明所要求的，这也解释了你那神秘的魔法师（或者当你假定自己是这个角色的时候）

是如何发现哪张牌被抽取了的。

对于表5-5所示的顺序，我们想象把方片A（原序列第1位）从卡牌的顶部移出并随机放置在卡牌中间的某处。比方说，方片A最终在方片4和方片10之间（在新分布的第4位和第10位之间）。从现在开始，它的顺序永久是错乱的，再怎么洗牌都不可能将其移动至原属的位置。



如果我们把一副扑克牌看做是一个无限循环，那么洗牌过程中的切牌不会影响整个序列。但是，非标准洗牌，比如将卡牌三等分切分，并在洗牌之前改变这三等分的顺序将会破坏序列。神奇的伎俩说明必须明确表示，卡牌应一次分成两堆。

当然，如果分析现实生活中玩扑克牌时会发生什么，就必须考虑人的影响，毕竟是人就会犯错。正如哲学家说的那样，“洗牌糟糕的是人类”。在一次完美鸽尾洗牌中，有些卡牌本应正好被一张卡牌所分离，但也许有些不可预期的因素，使这些卡牌被两张卡牌所分离，或可能仍然相邻并没有被分离。表5-6显示了一个更人性化的、不完美洗牌的可能结果。

表5-6：马虎的洗牌对卡牌分布的可能影响

洗牌前	真实的人类鸽尾洗牌后
1. 方片A	1. 方片A
2. 方片2	7. 方片7
3. 方片3	8. 方片8
4. 方片4	2. 方片2
5. 方片5	3. 方片3
6. 方片6	9. 方片9
7. 方片7	10. 方片10
8. 方片8	5. 方片5
9. 方片9	4. 方片4
10. 方片10	11. 方片11
11. 方片11	6. 方片6
12. 方片12	12. 方片12

这种实际洗牌中的随机性，不但产生一个困境，还创造了一个机会。困境是，现在不能准确识别哪张卡是乱序的，因为该序列不能被完全重建，魔术师必须依靠一点概率，这就给伎俩增加了一些风险。

当这个伎俩的观众意识到你不可能实现完美洗牌时，机会出现了。当你在这种随机不确定中，不管以何种方式选出那张卡牌，观众的困惑都会更大。

5.4.2　成功的概率

因为不知道卡牌乱序的确切性质，魔法师能够识别出顺序错乱的那张卡牌只是因为洗牌不够完美。此外，如果一张牌从卡牌顶部取出后又放回到卡牌中间，此时指令不再允许切牌或洗牌，那么这个伎俩更容易成功（只有一张卡是乱序的）。

哥伦比亚大学和哈佛大学的统计学家戴夫·拜耳（Dave Bayer）和佩尔西·戴康尼斯（Persi Diaconis），按照这种神奇伎俩所描述的方式混合了一副扑克牌，对洗过的扑克牌的可能结果做了数学上的探索。（想必任教于这些机构的教员都有很多空闲时间？）他们为识别一张错位的卡牌而开发出了一个数学公式，并进行了一百万次电脑模拟测试他们的“网络巫师”所选卡牌的准确性。他们的分析假定是完美的燕尾洗牌。他们发现，只洗几次牌时，这个伎俩表现得相当不错，但是随着允许越来越多次数的洗牌，成功的几率迅速下降。

表5-7显示了对52张牌进行不同次数的洗牌时成功的概率，也展示了如果允许一次以上的猜测，正确的卡牌被选中的几率。

表5-7：看似不可能的成功几率

猜测次数	2次洗牌	3次洗牌	4次洗牌	5次洗牌	6次洗牌
1	99.7 %	83.9 %	28.8%	8.8%	4.2%
2	100 %	94.3 %	47.1 %	16.8%	8.3%
3	100 %	96.5 %	59.0 %	23.8%	12.3%

当然，当人们考虑现实世界洗牌的随机误差时，成功的几率会小幅下滑，但相对的成功率仍然如表5-7所示。如果你像描述的那样执行这个伎俩——3次洗牌后猜1次，那么你猜测正确的几率大约是80%（考虑到糟糕的洗牌，实际正确的几率比估计的83.9%低一点）。

为了确保这个伎俩的实施，你可能需要至少3个人。那么，假设每个人的可能性为80%，你会让这3人中至少一人惊奇的几率增加至98.4%，这几乎是一个必然。如果你3次都错了，那就别再对这些人说话或写信，关闭你的邮箱，并专注于生活中更重要的事情。毕竟，如果辛勤工作，未来某天你有可能会进入哥伦比亚大学或哈佛大学，做真正重要的东西。

5.4.3　参阅

- 拜耳和戴康尼斯的研究出现在1992年《应用概率年鉴》（*The Annals of Applied Probability*）的第2期，294~313页。在那篇文章里，他们引述了两位魔法师的研究成果，这两位都是研究**上升序列**原理卡牌技巧的早期开发人员（如下所示）：
- Williams, C.O. (1912).“A card reading.”*The Magician Monthly*, 8, 67.
- Jordan, C.T. (1916).“Long distance mind reading.”*The Sphinx*, 15, 57. 这是本Hack所描述效果的依据。



5.5 检查你 iPod 的诚实性

找出你的iPod“随机”打乱顺序的真正随机程度。

苹果公司的iTunes是允许你在iPod上播放歌曲的软件，其中个性化的歌曲评级可以让你迅速找到你的最爱，这有助于派对随机播放（Party Shuffle）功能更多地播放你最喜欢的歌曲。iTunes挑选播放列表里下一首歌时使用的算法是：从你的最爱里随机选择。但它是真的随机吗？

在iTunes里，如果你反复听到音乐库里一位艺术家的歌曲，你可能会认为你的播放器有它自己的偏好。不过苹果声称iTunes里的歌曲的打乱算法是完全随机的。打乱算法选择的歌是*无放回*的。也就是说，就像遍历一副洗好的扑克牌一样，在你听完所有歌之前，每首歌只会听到一次（或在你停止播放前，或选择不同的播放列表前）。

iTunes中的派对随机播放是另一回事。其算法选择的歌曲是*有放回*的，这意味着每首歌曲播放后整个音乐库被重新打乱（就像每次抽出一张牌后重新对整副牌再洗一次）。“较多播放评级较高的歌曲”选项的确做到了较多次地播放高评级的歌曲，但对高评级的歌曲有多少偏好？



本Hack最初在OmniNerd网站上http://www.omninerd.com/以一篇文章的形式出现。

5.5.1 评估iTunes的筛选过程

我想测试两种不同的歌曲选项：*派对随机播放*和“较多播放评级较高的歌曲“。我创建6首歌的短播放列表：5个星级中每个星级选一首，剩下的那首是没有评级的。这些歌曲属于同一流派和艺术家，并把每首歌的播放时长改为只有一秒钟。



我在iTunes 5版本上进行我的测试。iTunes 6增加了**智能随机播放**功能，这可能会降低连续听到同一艺术家或专辑的几率，但我没有测试它。

重置播放计数为零后，我点击播放按钮，然后离开我的办公桌去度周末。对这些歌我会播放两次：一次选择*随机*（派对随机播放），一次选择*随机*和“较多播放评级较高的歌曲”选项。表5-8显示了在星期一早晨的播放计数。

表5-8：歌曲选择的分布

	随机选择		基于评级	
歌曲评级	播放次数	百分比	播放次数	百分比
没有评级	9105	16.70%	2052	3.9%
1	9055	16.60%	6238	11.8%

5

（续）

	随机选择		基于评级	
歌曲评级	播放次数	百分比	播放次数	百分比
2	9090	16.67%	8125	15.4%
3	9114	16.71%	10 020	18.9%
4	9027	16.55%	12 158	23.0%
5	9146	16.77%	14 293	27.0%
总计	54 537	100%	52 886	100%

在随机试验中，所有歌的播放次数都非常接近，和随机选择预期的一样。对于基于歌曲评级（或评价偏向选择）的试验，偏好算法对评级歌曲12%~27%的选择几率似乎是线性的。从5星评级一直下降，星级评级每下降一级，线性偏好就下降4%左右，但从一星级到无评级，降幅加倍了，有8%的下降。虽然一星级似乎是最低的等级，但没有评级才是真正的害群之马。



你的iPod假定：如果你对一首歌没有进行星级评定，那么相比那些你给最低评级的歌曲，你更不想听到这些没有评级的歌曲。这有点像选择一部差评的电影，而不选择暂时没有任何评价的电影。

图5-2显示了不同的歌曲选择选项的效果。你可以通过观察图表上的"随机"条的高度，来判断真正的随机选择选项的随机性。"评级偏差"条的线性性质，可以通过分析从1星级到5星级的每一步移动是否有相同的上升高度来判断。

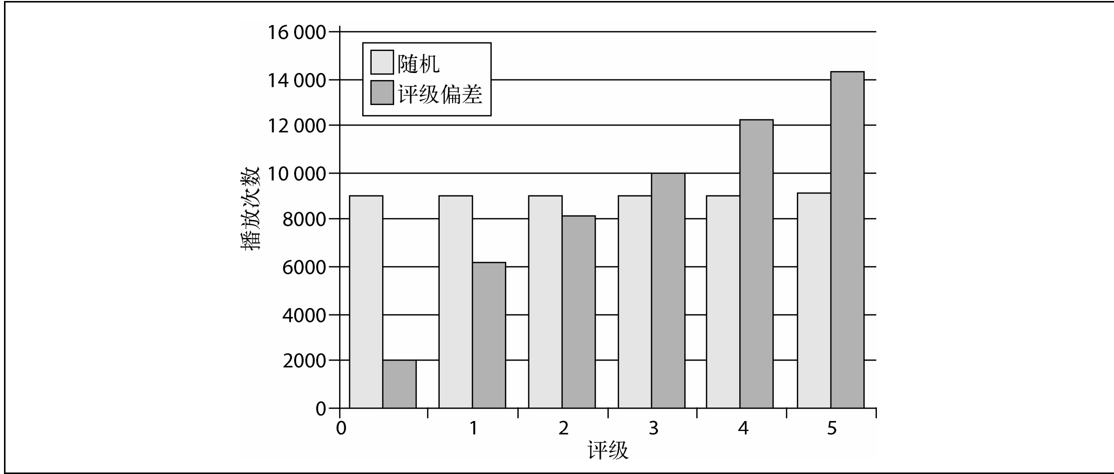


图5-2：歌曲选择模式

5.5.2　计算选择过程的统计量

改变每个评级内歌曲的数目会改变每首歌曲被选中的概率。因为每个评级里都有多首歌曲，

一首被评为r的歌曲在下一次评级偏好的（ratings-biased）派对随机播放中出现的几率，可用如下表达式计算：

$$\frac{x_r P_r}{(x_0 P_0)+(x_1 P_1)+(x_2 P_2)+(x_3 P_3)+(x_4 P_4)+(x_5 P_5)}$$

这个表达式中的下标表示歌曲的评级。一首歌曲被选中的几率取决于x（每个等级的歌曲数）和P（iTunes算法给的每个评级的百分比权重）。

每类评级的iTunes偏好概率，取自一个周末的抽样，下面是结果表达式：

$$\frac{x_r P_r}{0.0388x_0 + 0.1180x_1 + 0.1536x_2 + 0.1895x_3 + 0.2299x_4 + 0.2703x_5}$$

虽然评级较高的歌曲会优先考虑，但相比其他所有的歌曲，你不一定会听到更多的5星评级歌曲。我们假设，多数人评级的时候遵循正态分布[Hack #23]，其中3星级最常见。表5-9显示一个假设的、评级歌曲计数为钟形曲线的iTunes资料库。

表5-9：型的歌曲评级分布

歌曲评级	歌曲数量
没有评级	72
1	321
2	1527
3	1812
4	507
5	95

如果用我们的频率方程运行这些假设的数字，会得到如图5-3所示的分布。

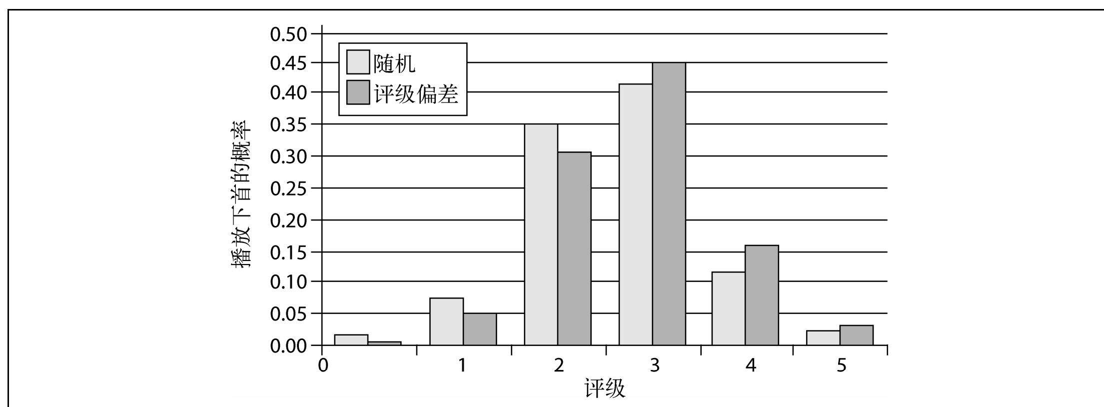


图5-3：歌曲选择的概率分布

正如你在图5-3中看到的那样，有特定评级的一首歌在播放列表中下一个出现的几率很大程度是由该歌曲评级内的歌曲数量决定的。iTunes对高评级歌曲的偏好，以及对低评级歌曲的厌恶只是略微提高或降低了由歌曲数量决定的概率。

可以将听到具有某个评级歌曲的几率运用到计算听到一首特定歌曲的几率。如果我们在歌曲选择表达式的分母里拿掉歌曲数量，我们就可以计算下一首是某首特定歌曲的几率，而不是下一首是某一评级歌曲的几率：

$$\frac{x_r P_r}{(x_0 P_0)+(x_1 P_1)+(x_2 P_2)+(x_3 P_3)+(x_4 P_4)+(x_5 P_5)}$$

5.5.3　解释统计惊喜

运行这些测试后大约一个月，我工作时发现我的iTunes派对随机播放时对同一首歌连续播放了两次。这是我第一次注意到一个连续的重复，然后我检查了播放列表。我不仅发现Nirvana的*Territorial Pissings*在列表上连续出现两次，而且A.F. I的*Death of Seasons*在三次音轨后连续出现了两次。

我用的是“较多播放评级较高的歌曲”这一选项，但这些都是中庸的3星级歌曲，我的歌曲库拥有近4000首歌。初看概率似乎令人惊讶，但你必须意识到你在一个工作日能听多少首歌。如果我平均每天工作10小时，而平均一首歌曲的播放时间是3.5分钟，概率认为我应该在不到一个月的时间内听完一个连贯的重复。

很多人声称当iTunes在漫步他们的音乐集时仍然可以看到模式（pattern），但这些模式大多数都只是同一个艺术家的多首歌曲。这么想想：如果你有2000首歌曲，40首来自同一个艺术家，随机播放时总是有大约2%的几率会再次听到这些歌。一首歌播放完后，同一个艺术家的歌曲在下35首歌再次播放的几率是50%，在下50首歌会再次播放的几率是64%。这可以通过下面这个公式来计算：

$$P(n)=1-\left(\frac{x_{\text{total}}-x_{\text{artist}}}{x_{\text{total}}}\right)^n$$

正如我们在其他Hack里看到的一样，小概率事件（比如我们有2%的几率能重复听到某位艺术家的歌曲）只需要一些机会[Hack #46]就会变成一个极有可能的事件。

我们会潜意识的找到一个模式，让你觉得iTunes有一个偏好。

5.5.4　参阅

有关iPod和洗牌的更多技术信息可以在下列资料中找到：

- Levy, Steven. “Does Your iPod Play Favorites.” January 31, 2005. http://msnbc.msn.com/id/6854309/site/newsweek/。
- Hofferth,Jerrod.“UsingPartyShuffleiniTunes.”August22,2004.http://ipodlounge.com/index.php/articles/comments/using-party-shuffle-in-itunes/.

——布雷恩·汉森



5.6 预测比赛冠军

借助提供的相关信息，我们有可能预测任何结果，尤其是体育领域的结果。有了多元回归技术和一个小软件，你就可以在比赛开始前猜测谁是赢家。诀窍在于选择正确的预测变量。

对相关性[Hack #11]的常见用法是找出两个变量共享信息的程度，或者更专业点，是两个变量之间方差的共享程度。



共享方差是一个数学术语，用来描述两个变量反映的冗余信息量。当大量的方差被共享时，预测是容易且准确的，因为对一个变量的认知就能导致对第二个变量的认知。共享方差通过对相关性进行平方来估计。

但是，我们的日常世界不止由单一变量预测另一个变量组成。事实上，在大多数情况下，用于预测特定结果的变量存在几个或多个。在这里，我们不处理单一变量对另外一个变量的预测，而是处理多个变量对一个变量的预测。这种工具称为多元回归（因为有多个预测变量）。

资深的体育赌徒、博彩公司和赌场运营者都熟悉多元回归，或者至少他们应该熟悉。有如此多的关于体育队伍的信息，以至于我们几乎能确定所有的变量，按照正确的组合方式，我们就可以相当准确地预测哪支球队会获胜。

投注职业足球是所有的赌博行为（至少我听到的是这样）中最常见的一种。这个技巧展示了如何收集数据，并使用多元回归预测足球赛的冠军。本例预测的是谁会赢得超级杯——全国足球联赛的冠军。

5.6.1 选择预测变量

第一步是构建模型（预测因子及其权重，你会用它们来进行预测）。对于足球，有很多关于球队过往成绩和球员特点的保存资料和统计数据。有些用来预测未来表现（例如，过往成绩）是合理的，而有些则不合理（例如，吉祥物的可爱度）。但是，赢钱的机会，是一个强大的动力，所以我会花时间和精力来收集所有我能收集到的关于每支球队和每一场比赛的统计资料。关键是

找到与赢得超级杯非常相关的变量。

我们假设你已经做完了相关研究，发现有6个变量与球队输赢有关。有些变量是合理的，有些不是合理的。你对获得最准确的真实生活预测感兴趣，所以甚至愿意将厨房水槽包含进去，如果它起作用的话。说明确点，你记录特定一支球队出现在超级杯中的年份，然后收集从那年起那支球队的数据。

想象一下，你已经发现，依据往年成绩和30支球队的特征，以下你感兴趣的变量可能在结果预测上是有用的。你在模型中使用的变量，以感兴趣的结果开始，也就是，在数据收集的那年球队是否赢得超级杯（是=1，否=2）？

你发现下面的变量和结果相关：

- 赛季期间轻松获胜的数量（超过9分）；
- 本赛季的平均出场数；
- 每场出售热狗的平均数；
- 团队佳得乐饮料的平均温度；
- 防守线球员的平均体重。

当你以真实的数据进行分析时，你可能会发现不同的潜在预测搭配。

5.6.2　将数据输入电子表格

社会科学家经常使用统计软件，如SPSS或SAS，但在这个例子中，我使用Excel工作表以及Excel非常酷的数据分析工具包（和回归工具）。我输入了一些虚构但符合实际的数据到表5-10所示的电子表格中。



什么？你以为我会告诉你一个预测足球比赛结果的真正秘密公式？我只是向你展示如何制作你自己的预测公式。我会自己留着的，非常感谢你！

表5-10：超级杯预测变量

队伍	是否赢得超级杯	轻松获胜次数	出场数	热狗数	佳得乐	体重
A	1	11	56 533	4798	56	276
B	2	9	44 543	5715	76	311
C	1	8	45 543	9753	45	315
D	1	6	45 768	8020	46	311
E	1	8	76 786	5395	56	256

（续）

队伍	是否赢得超级杯	轻松获胜次数	出场数	热狗数	佳得乐	体重
F	1	11	56 533	1054	67	277
G	2	9	56 554	750	76	256
H	2	12	44 675	6576	77	254
I	2	11	56 667	9187	77	287
J	2	10	65 545	4533	87	301
K	2	12	78 756	1963	86	243

表5-10显示了我收集的虚构的30行数据的一部分，30行数据代表我统计分析中用到的30个例子。数据的行数越多，你可以获取的例子越多，最终的预测也会越准确。

5.6.3 建立回归方程

你或许还记得高中时代的一个公式，简单的直线公式看起来像这样：

$$Y' = bX + a$$

这个方程由以下变量组成：

Y' 变量Y上的预测分数

b 该直线的斜率

X 分数的单一预测源

a 截距（直线穿越Y或垂直轴的地方）

因此，举例来说，如果你想用体重预测人类的高度，可以通过一组数据得出各个值，然后创建公式，你可能会得到看起来像这样的东西：

$$Y=35X+20.3$$

这意味着，如果你的体重（X变量）是125英镑，预测结果就是你大约高64英寸，或大约高5英尺3英寸。

但是，当我们有多个预测变量时，事情变得更有趣了。我们有了一个较长的系列预测（多个X）和权重（多个b）。

我在SPSS统计软件里使用该数据运行多元回归分析，你也可以使用Excel得到大部分相同的信息（见补充内容“在Excel中获得回归信息”）。

在Excel中获得回归信息

有两种方法可使用Excel来获得统计回归信息。首先，你可以使用SLOPE和INTERCEPT函数，你可以从Insert-Function找到。选择函数并输入参数（数据所在的单元格），Excel返回这些值，它允许你插入已知的值并预测其他的值。此方法在只有一个预测变量时效果最好。

你也可以使用数据分析工具库中的Regression选项，这是一个Excel加载项（你可能需要安装）。使用工具菜单上的这个选项时，你可以采用F检验测试回归系数的显著性，F检验类似于t统计检验[Hack #17]。

结果（即输出）如表5-11和表5-12所示。让我们看看哪个变量能最好地协助我们预测一支球队是否会赢得超级杯。

表5-11：回归统计

多元R	R^2	观察量
0.8483	0.7196	30

表5-12：回归方程

变量	系数	t统计值	P值
截距	−0.784	−1.010	0.323
轻松获胜	0.119	4.274	0.000
出场	0.000	−0.822	0.416
卖出热狗	0.000	1.043	0.308
佳得乐	0.013	2.457	0.022
体重	0.001	0.580	0.567

表5-12显示了方程的5个变量的系数（权重），用于表示每一个预测超级杯赢家的变量的表现情况。例如，和“轻松获胜”这一变量的相关系数是0.119。

如果我们将所有这些信息结合进一个大的方程来预测超级杯的结果，能得到如下模型：

$$Y' = bX_1 + bX_2 + bX_3 + bX_4 + bX_5 + a$$

所以，每一个预测变量（从 X_1 到 X_5 ）都有对应的具体权重（式中的*b*或结果中的系数）。

现在，把单词代入相同的公式：

$$b \times \text{获胜} + b \times \text{平均出场} + b \times \text{热狗} + b \times \text{温度} + b \times \text{体重} + a$$

使用表5-12所示输出的数据，下面是真正的实况回归方程：

$$Y' = 0.119X_1 + 0.000X_2 + 0.000X_3 + 0.013X_4 + 0.001X_5 + a$$

5.6.4 解释和运用回归方程

试想一下，对所有输入到电子表格中的行数据使用这个方程。超级杯的实际结果和预测结果有相当高的相关性。我知道这是因为表5-11显示输出的“多元R”部分，显示了相当高的相关性。0.84接近于1，这是你能得到的最高的相关。



“R^2”为0.72，这就是我们之前谈到的**共享方差**的比例。

这是什么意思？这些预测变量的组合是判断一支球队是否会赢得超级杯的相当有效的方法。万无一失吗？当然不是，因为这些组合变量并没有完美地预测结果，但它确实做了一个非常扎实的工作。

那么，举例说今年Denver Cannonball的数据点如表5-13所示。

表5-13：Denver Cannonball数据

变　量	值
轻松获胜	13
出场数	35 678
热狗	4567
佳得乐	65
体重	267

将这些数据插入前面所示的公式，下面就是我们得到的关于Y的预测：

$$Y' = 0.119(13) + 0.000(35678) + 0.000(4567) + 0.013(65) + 0.001(267) - 0.784$$

Y的最终值是1.875，更接近2（意味着没有预测出他们会夺冠）而不是1（意味着他们预计将获胜）。

一套好的预测指标有哪些关键点？

- 所有的预测都应该是相互独立的（如果可能的话要完全独立），因为你希望在对预测的理解上，它们能提供独特的贡献。
- 每个预测变量应该尽可能高地和你预测的结果相关。

5.6.5 改进你的回归方程

仔细研究这个Hack产生的方程，可发现大部分的预测能力只来自于两个变量：轻松获胜的数量和球队的佳得乐温度。另外，许多预测变量的权重为零，这意味着你不需要它们。你可以删除

这些无用的变量（出场数和出售的热狗数）以简化你的公式。事实上，只收集轻松获胜数和佳得乐温度数据就足以在我们的例子中作出相当准确的预测。

——尼尔·萨尔金德



5.7　预测棒球比赛的胜负

打开你的收音机，停在棒球比赛电台5秒钟，然后将其关闭。不需要听到分数，你就可以说出获胜的一方，你有超过一半的次数是对的。

你看，我是个大忙人。我一直在寻找一种方式节省花在生活中不太重要的事情上的时间，比如追随我喜爱的本地棒球队，这样我就有更多的时间花在生活中重要的事情上：朋友、家庭、讨论Holm’s sequential Bonferroni方法[3]作为方差分析的合适补充方法，等等。一个典型的例子就发生在几天前。我想知道堪萨斯城皇家队是否会赢得一场正在进行的棒球比赛，但我几乎没有时间等到比赛结束。我现在就想知道结果！



就像维鲁卡·索尔特[4]和她对拥有一个威利·旺卡[5]工厂里的奥古伦伯人[6]的兴趣一样，“就现在！”，我没有太多的耐心。

就像一个晴天霹雳，我意识到，我可以打开车上的收音机，只需短短几秒钟，我就能有足够的信息来猜测比赛结果。我能做到这一点，且不需要听得分情况或谁在垒上。

5.7.1　如何生效

在棒球比赛开始后的几个小时内，打开那场比赛的广播。收听时长能刚好确认哪个球队在击球即可。那支球队有大于50%的几率赢得比赛。

5.7.2　生效原理

棒球是这样一种比赛：你进攻的时间越长，你能获得的分数越多。随着一局中出现越来越多的击球员，沿着垒径跑动的击跑员和穿过本垒板的几率增加。另一种看待它的方式是，想象一局比赛快结束时，某支队伍获得了很高的得分。如果这个球队得分很多，他们必定已经使用了比那局规定最少的3个击球员更多的击球员，因此，在垒上的时间比其他的球队要长，长出的时间和队员数成比例。比赛过程中，在垒上时间最长的球队更容易得分多（或有更多的成果显著的赛局）。

注3：Holm’s sequential Bonferroni方法是在统计学中用来控制I型错误的方法。——译者注

注4：Veruca Salt，《查理和巧克力工厂》中的人物。——译者注

注5：Willy Wonka，是《查理和巧克力工厂》里的一名虚拟角色。——译者注

注6：Oompa-Loompas，奥古伦伯的本地人，作为工厂中的小矮人工人，他们表演了一场又一场精彩的舞台剧。——译者注

抽样理论[Hack #19]表明，样本最有可能捕捉总体中最常见的元素。在这里我们的总体是一场比赛中我们能听到的所有时刻。总体中最常见的特征（用“谁在垒上”表示）属于在垒上时间最长的球队。

图5-4显示出常规9局比赛的垒上时间的可能分布。在这个例子中，获胜的球队有58%的时间处于进攻状态。现在回想起来，随机找个时间打开广播，有58%的几率发现获胜的球队在垒上。

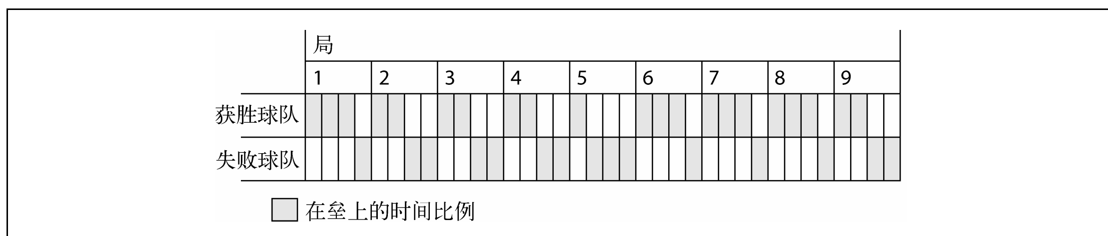


图5-4：输赢球队的垒上时间

从长远来看，利用棒球广播节目信息预测比赛结果的准确度应高于50%，但不一定真的很准。这是因为在垒上的时间和进球获胜之间的关系不是一个完美的相关[Hack #11]。球员可以得分快点，比如，在他们的第一球击中一个本垒打；或者他们可以花自己的时间获得很多击打次数，但困住了很多击跑员，从不得分。

但是，总体上，两个变量之间的相关性应是正的。即便图5-4中我想象的数据可能仅有58%的准确率，但这也比你盲目猜测的准确率高出16%。如果在21点牌桌边有这样一个优势，你会在一个星期内变成百万富翁。

5.7.3　证明有效性

为了测试我观点的正确性，你可以使用日报上出现的数据。虽然大多数比赛记录表没有每支球队垒上总时间这一信息，但有一个变量提供了几乎相同的信息。记录表肯定会报告一个“总垒上数”（total at-bats）。虽然这个统计量和垒上时间不一样，但它们之间应该有相当高的相关性。每一天都会提供十几场比赛的这个信息，短短几天的有价值的数据应该足以检验我的理论。收集每支球队的总垒上数，包括哪支球队赢得了比赛。



现实生活中的研究人员往往无法获得他们真正想知道的变量，我们使用**垒上数量**代替**垒上时间**就是一个很好的例子。相反，我们必须选择另一个可用的最好的变量。科学家们称这些替代品为**代理变量**或**替代变量**。

我的假设是，垒上数最多的球队，赢得比赛的几率大于50%。出于好奇，我测试了自己的这

个假设。我用芝加哥小熊队作为例子,因为他们的统计资料在网上都是现成的。我随机选取了2003年小熊队最初的25场比赛。通过对这样比赛的分析发现，垒上数最多的球队，赢的几率为56%。如果我消除在垒上平局的3种情况，我预测的准确度就有63%。虽然垒上数最少的球队，有时也会战胜芝加哥小熊队，但垒上数的差距越大，垒上数最多的球队越有可能赢。当垒上数最多的球队获胜时，他们平均比输掉的球队垒上数多出4.14。当垒上数最少的球队获胜时，他们平均只比输掉的球队垒上数多出2.88。

5.7.4　其他生效领域

有人建议，当我支持的堪萨斯城皇家队参赛时，如果我想超过一半的时间是正确的，我应该总是预测他们输掉。是的，是的，这很搞笑。

5.7.5　无效领域

我建议你在赛程的前几个小时尝试，因为如果你在第九局打开收音机，那么这个方法的准确度会变低，根据棒球规则，如果主场队在第九局前处于领先，他们就不用击球。他们赢了。比赛结束。因为主场队往往比客场队赢得更多，这意味着获胜的球队常常在第九局从不击球。

这就提出了这种预测方法的一个有趣的变化，只适用于第九局。比赛进行到第九局时打开广播，如果你支持的球队正在击球，或许这并不是一件好事。芝加哥小熊队的数据显示，获胜的球队偶尔比他们的对手有更少的垒上数，这可以通过这样的事实来进行部分解释：获胜的球队有时只在前八局里击球。

这种方法并不适用于所有的体育运动。例如，在篮球比赛中，人们不认为持球时间和得分正相关，在激烈比赛中，快速进球的球队甚至导致相关性为负。另一方面，在足球场上，持球时间被认为是一个关键的能力表现指标，通常和胜利相关。



5.8　在 Excel 中绘制直方图

使用Microsoft Excel来绘制数据分布，可以让你对统计数据有一个更好的理解。

俗话说“一图胜千言。”这有一定的道理。一幅图往往是理解1000个数字的最好方式。人是视觉导向的。我们善于看一张图片并观察不同的特征，不善于看有1000个数字的列表。

直方图是帮助我们理解数据的一种最有力的工具，它是关于值分布的图。下面是直方图的概念。假设你有很多数据，比方说，1955年至2004年间，所有6032名每场比赛中安打数为3.1及以

上的棒球选手的安打率[7]。我们同时假定，你想知道这些值是如何分布的。最低值是多少，最高值是多少？低的值是否比高的值更多？安打率是完全介于0~0.400的随机数，还是存在某种模式？

安打率可以有许多不同的值。1955年到2004年，6032名球员有合格的安打率，有1229个独特的值。您可以绘制每个独特安打率下的球员数（虽然我无法想象这个图是什么样子）。但我们并不真正关心每一个独特的值，例如，13名球员有0.2 862的安打率不是那么有趣。相反，我们可能会想知道有非常相似安打率的球员数量，比如说安打率在0.285~0.290的球员数量是多少。

让我们把每个范围想象成一个桶。每个赛季球员进入一个桶里。例如，1959年，汉克・亚纶（Hank Aaron）有0.354的安打率，所以我们会把这个赛季放在0.350~0.355的桶里。所以，下面是我们的方案：我们把每个赛季球员放到一个桶里，计算每个桶里赛季球员的数量，并绘制图形展示（按升序排列）每个桶里球员的数量。这个图就是直方图。

5.8.1 代码

在这个例子中，我想看看安打率的分布。我用了包含每个球员每年总的安打统计的表格（以及每个球员所在的球队名单），还有我称为b_and_t的表格。我只选择了1955年至2004年间，获得足够打席数够格成为联赛冠军的球员：

```
SELECT  b.playerID , M.nameLast , M.nameFirst , b.yearID , b.teamG ,
b.teamIDs , b.AB , b.H ,
b.H / b.AB  AS  AVG ,
b.AB + b.BB + b.HBP + b.SF  as  PA
FROM b_and_t inner join Master M
on b.playerID = m.playerID
WHERE yearID > 1954
AND b.AB + b.BB + b.HBP + b.SF  >  b.teamG * 3.1 ;
```

运行此查询后，我把结果保存为Excel文件，名为batting_averages.xls。

在Excel中绘制直方图的一种方法是使用分析工具库（Analysis ToolPak）的加载项。你可以从Tools菜单通过选择Add…Ins来添加，然后选择分析工具库（Analysis ToolPak）。这给Tools菜单增加了一个新的菜单项，叫数据分析。它引入了一些新功能，包括直方图这个功能。但我觉得这个界面混乱而且缺乏灵活性，所以我用了别的方法。

下面是我创建直方图的方法。

(1) 在工作表的数据中，创建一个新的名为Range的列。

注7：在棒球运动中，安打表示击球手把投手投出来的球击出到界内，使打者本身能至少安全上到一垒的情形。
——译者注

(2) 在本列的第一个单元格，对你希望为其绘制分布的值使用函数进行四舍五入。做到这一点最简单的方法来是使用ROUND函数的有效数字选项。在我的工作表中，列I包含了我想计算分布的值（安打率），所以我可以用一个公式，比如ROUND(I2,2)来四舍五入到最接近0.010。就个人而言，我发现0.005大小的桶更具描述性，所以我用了一个技巧。你可以在ROUND函数里乘以一个值，然后在函数外除以一个值，这样可以得到几乎任何大小的桶。在ROUND函数里，我乘以桶大小的倒数——这种情况下，是1/0.005=200。函数外面，我乘以桶的大小。在我的工作表中，列I包含了平均值。于是，我用ROUND(I2*200,0)/200作为我的公式。将此公式复制粘贴到工作表中的每一行。（你可通过双击单元格的右下角，快速快做到这一点。）

(3) 现在，我们已经准备好计算每个桶中玩家的数量了。选择工作表的所有数据，包括新的Range列。从Data菜单中选择数据透视表和数据透视图表报告。选择数据透视图报表，然后单击Finish（我们将使用所有的默认值）。我们将为我们的数据透视表选择两个区域。从数据透视表字段列表面板中，选择Range。将这个拖放到数据透视表的Drop Row Fields Here部分。接下来，拖放“playerID”到数据透视表的Drop Data Item Here部分。默认情况下，Excel将计算匹配每个范围值的球员ID数。数据透视表现在显示了每个桶中的项目数。你应该看到一张（非常难看关于）的每个桶中球员数量的图。

(4) 清理美化图表。（我喜欢擦除背景填充和线条，改变列宽。）图5-5的例子即为一张清理干净后的图表。

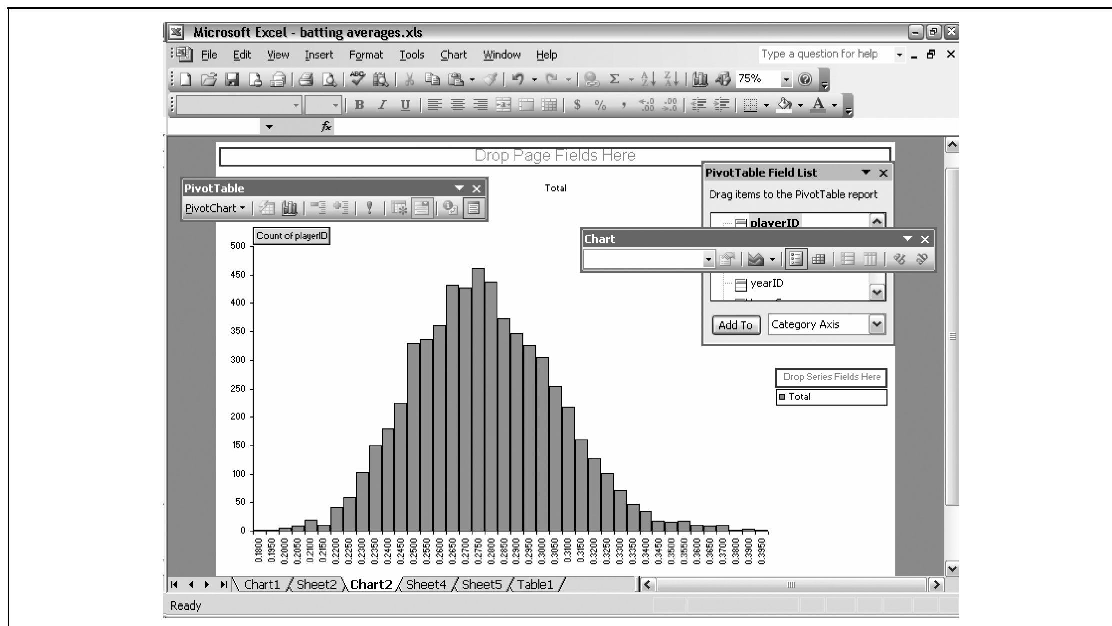


图5-5：数据透视图报表的直方图

通过直方图，我们看到，分布类似于一个钟形曲线，它向右侧倾斜，中心大约在0.275左右。

5.8.2 解读Hack

用公式计算组条的好处之一是，你能很容易地改变组条的公式。以下是采用其他公式时的几点建议：

ROUNDDOWN（*<value>*，*<significance>*）和ROUNDUP（*<value>*，*<significance>*）

这个ROUNDDOWN函数向下舍入到最近的有效数字。例如，ROUNDDOWN(3.59,0)等于3，ROUNDDOWN(3.59,1)等于3.5。同样，ROUNDUP是向上舍入到最近的有效数字。ROUNDUP(3.59,0)等于4，ROUNDUP(3.59,1)等于3.6。

LOG（*<value>*，*<base>*）

使用对数大小的组条时，有时可用此函数绘制对数刻度值。您可以结合LOG函数和ROUND函数来创建可变大小的组条。

CONCATENATE（...）

CONCATENATE函数不计算数字，它把文字放在一起。如果你想明确列出范围（如3.500~3.599），可以使用CONCATENATE来创建。例如CONCATENATE(ROUNDDOWN(3.59,1),"to",ROUNDUP(3.59,1)-0.01)返回3.5到3.59。

如果你想更进一步，可以用一个命名值取代组条。（例如，命名单元格A1 bin_size）。这可以很容易地动态改变组条大小和试验不同数量的组条。

——约瑟夫·阿德勒



5.9 去得两分

在橄榄球比赛中，什么时候尝试两分转换是正确的选择？无论你使用哪个“图表”，当统计学家都对此有争议时，问题变得更加复杂。

几年前，当我的本地职业橄榄球队正在输掉一场势均力敌的比赛时，我正非常享受地观看他们的比赛。与其说我对本地球队的低迷表现感兴趣，不如说我对迷糊的教练试图读懂两分转换图表感兴趣。



在橄榄球中，触地得分后（触地本身得6分），得分队有两种选择：获得一个“加分”或两个加分。通常情况下，得分队会选择通过将球踢入达阵区底部架设的两根球门柱而获得一个额外的分数（像短距离射门得分一样），但他们也可以选择“达阵再达阵”分（称为两分转换），指进攻方以跑或传的方式再次通过对方的达阵区底部。

后来体育新闻记者“证实”，很明显，当时教练不知道如何读图表。具体来说，当他解读代表球队落后或领先多少分的那列时，他认为那意味着是他们得到转换分数之后，一支球队落后或领先的分数。

正当我沉思为什么一个国家橄榄球联赛（NFL）主教练没有学会如何读懂图表时，我开始想知道是谁制作了这张“图表”，它是基于什么原则制作的。后来，我搜索了“官方图表”，我发现两张“官方”图表并不完全一致。

最近，我碰到一张基于可能结果的概率统计分析和剩余时间（由剩余的持球数表示）的图表。这张图表和我之前发现的图表都不一样。

这个Hack是写给你的，教练。它从统计学的角度分析，什么时候该去得两分，什么时候应该满足于再得一分。

5.9.1　传统的两分转换图表

当你在电视上看到一个教练拿着一张塑料夹层卡，在决定是否得两分前研究它时，体育节目解说员喜欢把这张卡称为图表，不过，正如上一节说到的那样，有很多可以使用的图表。导致这种细微差异的原因可能在于一种图表被国家橄榄球联盟（NFL）采用，其他的作为经典标准决策集被大学橄榄球比赛采用。

差异也可能基于这样的事实：大学的图表是为特定的、可能更积极或更自信的球队制订。大学图表似乎为胜利而制订，而不是平局。虽然大学橄榄球现在有加时赛的规则，但还处于起步阶段，而专业橄榄球比赛有加时赛已经很长时间了。

国家橄榄球联赛（NFL）的图表由诺姆·希茨格斯（Norm Hitzges）的网站http://www.normhitzges.com/thechart.htm提供，（诺姆是达拉斯的一位播音员，还是一个全能运动专家）。大学图表（在http://www.NFL.com/fans/twopointconv.html）正式使用于20世纪70年代，由加州大学洛杉矶分校（UCLA）开发制得。表5-14对两张图表的决策建议进行了整合。

表5-14：两分尝试的经典决策

	分数落后或超前												
	0	1	2	3	4	5	6	7	8	9	10	11	12
落后（NFL）	1	1	2	1	1	2	1	1	1	1	2	1	1
落后（大学）		2	2	1		2	1	1	1	2	1	2	2
	0	1	2	3	4	5	6	7	8	9	10	11	12
领先（NFL）	1	2	1	1	2	2	1	1	1	1	1	2	2
领先（大学）		2	1	1	2	2	1	1	1	1	1	1	2

加州大学洛杉矶分校的图表没有提供当比分为平局时的决策建议，也没有提供当你的球队处于落后时的建议。另一方面，NFL的图表对所有场合和情形都提出了决策建议。如讨论的那样，主要差别似乎在于你是否愿意为平局而战。加州大学洛杉矶分校显然不想为平局而战，而NFL的图表则没有这样的顾虑。

5.9.2　现代超级科技图表

在现实世界中，一组统计概率控制了体育赛事的结果，关于是否去得两分或获得额外一分的决定，应根据更多的信息来判断，而不仅仅根据分数和球队的输赢信息。在实际比赛的情况下，聪明的教练会将如下额外因素考虑进去：

- 射门球员射门得分的可能性；
- 球队在给定的两分转换情况下得分的可能性；
- 球员目前的健康状况、态度和技能；
- 球队还会获得多少持球机会。

过去的统计数据显示，平均而言，国家橄榄球联盟（NFL）的橄榄球队，获得额外一分的几率大约是98%，获得两分尝试的几率约为40%。教练必须利用自己的经验和直觉来衡量自身球员目前的能力水平，而这个分数和图表没有任何关系。

然而，对于剩余的持球机会，这正是建立在概率基础上的决策系统需要考虑的信息类型。从结束点回推假设出来的、考虑了两种选择概率（98%的一分和40%的两分）的橄榄球比赛，统计学家们已经制作出不仅基于目前得分，同时也基于两支球队剩余持球机会总数的图表。

在2000年《几率》（*Chance*）期刊（第13卷，第3号）中，哈罗德·萨克罗维茨（Harold Sackrowitz）使用*动态编程方法*提出了新的分析结果。表5-15显示了萨克罗维茨博士所做图表的一部分。

表5-15：两分尝试的现代决策

落后或领先的分数														
		0	1	2	3	4	5	6	7	8	9	10	11	12
剩余持球数														
1	落后	1	1	2					1					
	领先	1	2	1	1		2	1	1	1				
2	落后	1	1	2	1	1	2		1	2		2		
	领先	1	2	1	1	1	2	1	1	1				
3	落后	1	1	2	1	1	2		1	2		2		
	领先	1	2	1	1	1	2	1	1	1	1	1	1	2

（续）

落后或领先的分数														
4	落后	1	1	2	1	1	2	1	1	2	2	2	1	
	领先	1	2	2	1	1	2	1	1	1	1	1	1	2
5	落后	1	1	2	1	1	2	1	1	2	2	2	1	
	领先	1	2	1	1	1	2	1	1	1	1	1	1	2
6	落后	1	1	2	1	1	2	1	1	2	2	2	1	2
	领先	1	2	2	1	1	2	1	1	1	1	1	1	2

这个两分转换图表基于比赛中所有可能的起始分数，假设了额外一分或两分转换情况下成功的基本概率。一个普通国家橄榄球联盟（NFL）比赛小节，总共有6次持球进攻的机会，所以把这张图表视作在第四小节最有用。萨克罗维茨还假定加时赛50%的几率获胜。

5.9.3　如何生效

表5-15的计算原理和下面的简单示例一样：

(1) 想象一下，你落后一分，且再次得到球的可能性不大。

(2) 你有98%的几率通过射门获得额外一分，你也有50%的几率在加时赛中获胜。获得额外一分导致你有49%的获胜几率（0.98 × 0.50=0.49）。

(3) 你有40%的几率进行两分转换，所以达阵再达阵获得两分导致你有40%的获胜几率。失败结束比赛，成则功赢得比赛。

(4) 49%比40%更好，所以你应该选择额外一分。请注意，如果你相信你的球队两分转换的几率比49%更高，你应该选择两分。按照这样的计算思路，经过一个较长的连续持球，就形成了表5-15所示的决策树。

下一次你指导至关重要的橄榄球比赛且需要作出关键决策时，你应该使用哪张图表？这取决于你自己，但要记住那位迷糊的橄榄球教练，那个几年前我在电视上看到的那位，他在次年被Dick Vermeil取代。Dick Vermeil被认为是更聪明的教练，而且帮助开发了加州大学洛杉矶分校的两分转换表，如表5-14所示。现在你知道这个故事的剩余内容了！



5.10　按优劣程度排序

有很多方法可以使用数据判断任何体育项目上谁是最好的。然而，所有比较体育运动上的个体表现的直观方法都需要考虑效度问题。

我和我的朋友们总是在竞争。最近一段时间，我们的战斗舞台一直是扑克。按照惯例，我和朋友们聚集在我家，参加一个得州扑克锦标赛。这是一个非正式的比赛，但大家都对它非常认真。

我们的扑克锦标赛的规则是：每个人开始都用等量的筹码，当筹码没了，他们也就离开了。有一人第一个离开，有一人最后一个离开，还有一些人在中间离开。因此，举例来说，如果7个人打比赛，就有人排在第一、第二、第三、第四、第五、第六和第七。

我们都认为自己技艺不错且相当有竞争力，我们渴望有一种客观的方法来比较比赛表现。作为该组中的统计学家之一，我当仁不让地设计出具有某种客观指标的多种方法，这种指标使得所有参赛者能够将他们之间的表现进行相互比较，从而能一劳永逸地判断谁是最好的玩家，谁只是偶尔交好运而已。这是关于我探索和选择统计解决方案的故事。我并不是要把结果说出来，但我知道，没有一个统一的最佳解决方案。

5.10.1 如何公平排序

一些有竞争力的组织，如体育联盟和协会，经常遇到如何确定最佳这种问题。问题的关键是如何在各种类别、场地和场合概括总结表现。

在体育界，有3种常用的方法可用于作出谁“最好”的判断。所有的方法在直觉上都基本讲得通，但每种方法都有其特定的优缺点。

首先，让我们来看看，我要分析的数据的性质。你的数据可能和我的数据类似，不论你运行的是每周家庭大富翁游戏的数据还是职业高尔夫协会的数据。虽然扑克不是一项运动，但任何有组织的、有竞争的努力都能提供排名的数据。表5-16显示了我自己的夏季联赛的扑克比赛结果。

表5-16：夏季扑克联赛数据

	保罗	丽莎	比利	贝宁	马克	布鲁斯	凯茜	蒂姆	戴维
5/14	6	5	4	3	2	1			
5/21	3	6	4	5	7	2	1		
5/28			5	4	1	3	2		
6/4			4	6	3	7	2	5	1
6/11			4	5	6	1	2	3	
6/18			5	4	2	3	1		
6/25			1	4	3	5	2		
7/2			1	5	4	3	2		

你可以看到，9名玩家至少都参加了一场比赛，但没有任何一场是所有玩家都参加的。如果一个人在某一天晚上没有数字，那是因为他没参赛。这在体育运动中是常见情况，如高尔夫球和网球。

在两种情况下，7人上场，但在其他场合，只有5人坐在一起打牌。有4个人参加了全部的8场比赛。（这些都是铁杆玩家，他们不得不承认，他们在认识什么是生命中最重要的这件事上有点问题。）有一名玩家戴维，只参加了一场比赛。

玩家名字下的数字表示他们的出局顺序。如果有6名玩家，你第一个出局，那你会得到一个点数，排在最后一名。如果你是6名玩家的赢家，你会因为是第一名而得到6个点数。



这个计分系统有一些需要注意的地方。首先，你只要参赛就会得到至少一个点数。其次，如果有更多的玩家参与，你就需要更多的积分来赢得比赛。

那么，如何在扑克联赛中对玩家进行等级排序？以下是3种常见的解决方案，所有这些方案都多少起点作用。

1. 总点数

对我而言，首先浮现在脑海中的是简单地把各场比赛的点数加起来，并根据玩家的总点数对其进行排序。这是名人按收入排名或银行劫匪按自己的犯罪数量排名时采用的方法。只需要参加很多比赛就能提升你的名次。要想成为年度的高尔夫球手，你必须参加很多场比赛，此外在这些比赛中，你的表现要过得去。

2. 平均表现

第二种方法是用总积分除以玩家参加的比赛数量得到平均点数。产生一个平均点数的妙处在于，你得到了一个代表典型表现水平的数字。这对测量难以捉摸的东西是理想的，比如天赋。你在扑克中（或任何其他赛事）的平均表现应该是能力的最佳单一指标。

3. 总的获胜数

第三种方法在团队运动中最简单、最常用，即计算胜利的次数。最经常获胜的玩家是最好的玩家。此方法适用于锦标赛风格的扑克（我们玩的那种），以及任何有一个明确赢家的赛事。

5.10.2　比较 3 种方法

每种排名方法都有各自明显的优势，并各司其职。表5-17展示了在这3种排名系统下每个玩家的值。

表5-17：扑克表现摘要

	保罗	丽莎	比利	贝宁	马克	布鲁斯	凯茜	蒂姆	戴维
点数	9	11	28	36	28	25	12	8	1
平均点数	4.5	5.5	3.5	4.5	3.5	3.13	1.71	4.0	1.0
获胜次数	1	1	2	1	2	2	0	0	0

所有这3个评分系统都是合理的。但是关于谁是最好的问题，这3个系统都有不同的答案！对像我这样的扑克科学家来说，这无疑是一个令人沮丧的发现。因为这3种方法都有理由被认为是“最好”的排列方法，而每个方法都产生不同的“最佳”扑克玩家，这有点矛盾。表5-18展示了

采用各个计分方法的排名区别。

表5-18：扑克排行榜

	保罗	丽莎	比利	贝宁	马克	布鲁斯	凯茜	蒂姆	戴维
点数	7	6	2.5	1	2．5	4	5	8	9
平均点数	2.5	1	5.5	2.5	5.5	7	8	4	9
获胜次数	4	4	2	4	2	2	6	6	6

请注意每个系统下的“最佳玩家”有什么区别。在总点数系统下，贝宁是最好的；在平均表现系统下，丽莎是最好的；在总的获胜数系统下，3人并列第一，但贝宁和丽莎却不在其中。3种方法唯一真正的一致是：戴维被评为最差的玩家。（对不起，戴维，但数字不会说谎，我为公众的嘲笑感到遗憾。也许我可以把这本书的免费复印本送你，向你示好？）



我指派排名时，通过将那些平局的人进行平均来打破平局。换句话说，比利、马克和我自己在获胜数系统下并列排名第一，所以1、2、3的排名，平均后是2，这就是我们的排名。

如果有3个不同的评分系统产生3个不同的排名，很显然，它们不可能都同等有效。它们不能都以相同方式产生真正体现我们感兴趣变量的分数，这个变量定义为玩扑克的能力。解决方案没有包含最佳方法的选择。我的目标不是确定最佳的系统并采用它，我的目标是提供有效的信息，让别人按他们的需求解释他们的数据。

我的解决办法是提供基于3种计分方法的所有3种排名。这样一来，玩家可以选择把重点放在对他们最有意义的方法所产生的排名结果上。

5.10.3 故事的结尾

在我的扑克联赛中，对玩家最有意义系统是让他们排名最高的系统。想象一下。

“任何一种方法可能都是可接受的、准确的。”带着这个认识，晚上我安稳地睡着了。毕竟，这3种方法中没有一种，会犯这样一个错误：得出我是最好玩家。关于这些方法，其中或自身一定有某种效度证据！

现实生活中的职业体育组织通过创建复合的积分系统来处理单个系统的优缺点。一些在网球和高尔夫球赛中（还有扑克锦标赛）用来改善排名系统的措施包括：

- ❑ 结合很长一段时间的表现数据；
- ❑ 对赢得更困难的比赛给予更多的点数；
- ❑ 同时使用平均表现和总点数系统，以奖励优秀球员和频繁参与的球员。

有点讽刺的是，这些系统中可能更公平、更准确的系统常常被媒体和球迷认为过于复杂和疯狂。使排名系统更有效的尝试，经常被公众视为无效而遭到拒绝。



5.11　通过几率估计圆周率

统计学家认为任何重要的东西都可以使用统计数据来发现。这可能是正确的，因为事实证明你可以使用统计信息来估计科学中最重要的基础值之一：圆周率。

计算圆周率是所有崭露头角的天才的常规技能之一。比如，我记住的22除以7的结果就非常接近准确值。还有多种计算方法，其中一些比其他的更为精确。不过，我最喜欢的方法，是采用几率和漫长的、寂寞的海上航行或其他强制的孤独时间等元素。好奇吧？继续往下读吧。

在展示如何估计圆周率值之前，我将以介绍几个基本的几何事实来开始我们的讨论。不要恐慌，我对几何懂的不是很多，所以我们不会在这上面花很多时间。我只会大致讲解一些基础知识，使我们能了解这个技艺的魔力。

5.11.1　圆周率

在几何里，圆周率是一个值大概为3.141 59的数（用π这个符号表示），人们已经发现圆周率和圆形的不同部分之间的关键关系，如图5-6所示。

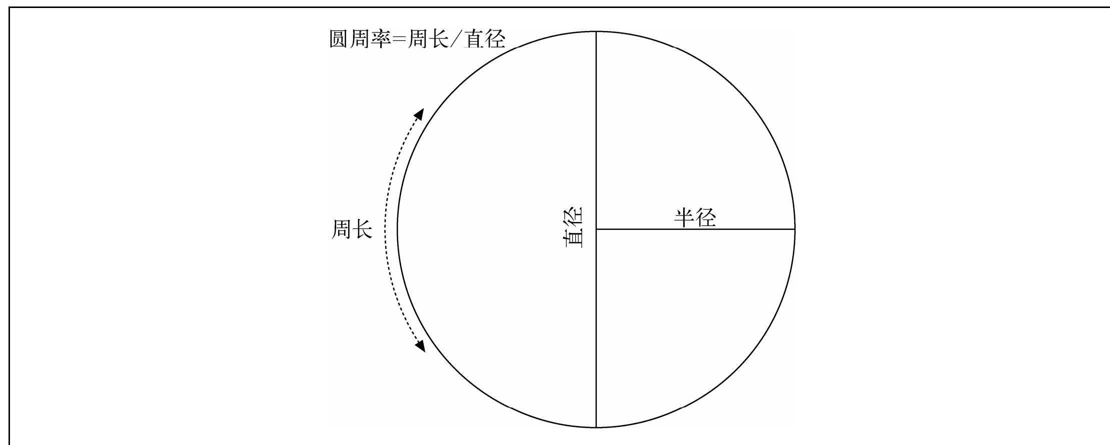


图5-6：计算圆周率

例如，如果你用圆的直径乘以圆周率，你会得到圆的周长。如果你把圆的半径平方，再乘以圆周率，你会得到圆的面积。

也许，这都很酷，但这是那些喜欢几何的人的主要兴趣，不是统计学家的主要兴趣。但只要等一等。

5.11.2　圆周率和落针

在1700年，乔治–路易·勒克莱尔（Georges-Louis Leclerc）向全世界提出了半几何/半统计的难题。他被称为布冯伯爵，或有诸如此类的称号，所以这个问题被称为布冯投针问题（Buffon's Needle Problem）。他提出了个大概，没有细节，我在这里总结一下。

想象一下，一个针头随机落在两条平行的水平线上。两条线之间的距离远大于针的长度。针落在其中一条线上的几率是多少？

有些问题你第一次听到时觉得不可能解决，这个问题就是其中之一，但它是可以解决的。没有必要在这里花费时间计算最终结果，但我肯定能做到这一点，我向你保证。真的，我可以。真的。解决方案涉及一些几何知识，它考虑到了两个关键的信息成分。任何给定的随机落点位置的关键在于：

- 针的中心距离最近的水平线多远；
- 针和最近水平线的垂直线的夹角是多少度。

用这两个信息定义针的随机位置形成了一些有助于简化问题的通用观察。

- 如果针的中心正好在一条水平线上，那么不管它的角度如何，针总是触碰到那条线。
- 如果针的中心足够接近水平线，距离小于针长度的一半，那么针有时会触碰到线。针的角度决定了针是否会触碰到线。
- 如果针的中心踞线的距离超过针一半的长度，那么不管针的角度如何，针永远不会触碰到线。
- 针越接近平行线，针碰到那根线的几率就越大。

所有可能的落针位置可以绘制为一条曲线，展示所有可能的与线的距离以及所有可能的与垂直线的角度。图上有三角函数，数学家们已经用下面的等式定义了这样的曲线：

$$\text{概率} = \frac{(2)(\text{针的长度})}{(\pi)(\text{线间距})}$$

这是问题的答案。让我们赶快用一些真实的数字试试吧，只是为了验证Leclerc的工作。想象一根3英寸长的针随机掉落在缝纫台上，台上有两根距离为4英寸的平行线。

针碰到其中一根平行线的几率是多少？以下是必要的计算：

$$\text{概率} = \frac{(2)(\text{针的长度})}{(\pi)(\text{线间距})} = \frac{(2)(3)}{(3.1459)(4)} = \frac{6}{12.566} = 0.477$$

针碰到其中一根线的几率约为48%。



当你想到一个地板上满是落针和线的大房间时，你的赌博之心是不是蠢蠢欲动了。去吧，给你更多的力量。这一法则已经在一些你可能见过的嘉年华游戏里生效了。有没有注意到，那些乒乓球落入鱼缸或者足球通过铁环的次数是多么稀少？

5.11.3 概率和圆周率

我保证你可以使用几率来估计圆周率，不过，不是使用圆周率来计算几率。数学的力量使得我们能够移动等式中的任何元素，所以等号右侧的任何元素符号都可以被移动到左侧。我们可以像这样移动我们的概率公式计算圆周率：

$$\text{Pi} = \frac{(2)(\text{针的长度})}{(\text{概率})(\text{线距间})}$$

我会使用我们测试这个概率等式时用的数字来证明其有效性。我们已经知道圆周率的正确答案是什么了，那么让我们来看看公式是否生效：

$$\frac{(2)(\text{针的长度})}{(\text{概率})(\text{线距间})} = \frac{(2)(3)}{(0.477)(4)} = 3.1447$$

该公式计算的圆周率的值为3.144 7，这相当接近3.141 59。如果我们允许我们的数字有更多的小数位，那我们可能会有一个更准确的答案。

5.11.4 使用概率估计圆周率

在我们的例子中，我们知道概率，所以我们可以利用这个信息来计算圆周率。但是，如果你不知道圆周率，却需要计算它的时候呢？如果你被困在一个荒岛上，或长时间在海上航行或一条腿骨折躺在床上，无法获得关于圆周率准确值的相关参考资料？进一步讲，假设你需要使用圆周率计算圆的周长或球的体积，或者计算几何、金融、物理中任何的值。一场噩梦，对吗？你可以使用这个公式，只需进行一场实验并收集数据，就能相当准确地计算圆周率。

用两条水平线设立一个区域，撒一些针，并保持跟踪。测量平行线间的距离以及针的长度，剩下的繁重工作就交给几率吧。从很多落针收集大量的数据样本，得到精确到小数点后几位的概率，也许约有一千次落针。祝你好运，继续认真记录。

比方说，你画两条距离为8英寸的平行线，使用约7英寸长的编织针。如果使用这样的设备进行了大量落针实验，你可能会发现针触碰到线的几率介于50%~60%。假如说，是55%。要使用这个数据来计算圆周率，你将像这样运用数学：

$$\frac{(2)(\text{针的长度})}{(\text{概率})(\text{线间距})} = \frac{(2)(7)}{(0.55)(8)} = \frac{14}{4.4} = 3.18$$

你会发现，3.18这个值非常接近图5-6所示的周长和直径之比。

如果你的视力不如从前，就没有必要使用难以看见的针。你可以使用相同的逻辑，让一支铅笔落在你的办公桌上，或将弹珠滚动到一个定义好的区域内，或让跳伞者降落在一个长方形的目标内。你需要两条平行的线，两条让铅笔、弹珠或跳伞者可以有机会降落进去的线，还有你需要知道物品的长度。只要结果是随机的，什么物品都可以，找到降落在草垛上的跳伞者比找到一根针更容易。

第 6 章

精明思考

（*Hack #61~#75*）

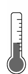
本章着重讲解那些可以帮你更清晰、更明智或更具创造性地进行思考的Hack。开篇我们利用概率法则，证明你自己比超级英雄更聪明[Hack #61]。通过掌握统计捷径[Hack #66]以及发现舞弊的能力[Hack #64]，我们能够长久地感觉自己很聪明。

随后，通过发挥你的怀疑面使你自己和他人印象深刻：揭秘惊人的巧合[Hack #62]，揭示怪异现象的真相[Hack #63]。反驳（或者证明）超能力（ESP）的存在[Hack #68]后，当你读懂你朋友的心思时，他们会感到惊讶[Hack #67]。

最后，学习如何规避常见的、不合逻辑的陷阱[Hack #69]，从而完成自我完善的课程。

既然你这么聪明，那么对你而言，注意到周围其他人没有注意到的事情应该是件轻而易举的事。你可以掌握交通拥堵的艺术[Hack #74]，探寻你和凯文·贝肯（Kevin Bacon）或是其他人之间的联系[Hack #72]，识别出只有政治科学家才知道的虚假的选举制度[Hack #73]。

本章的最后讲解能够扩展你视野的Hack。尝试不同的、令人兴奋的职业，比如间谍和密码分析[Hack #70]，发现新物种[Hack #71]，也许，甚至是其他星球的生物[Hack #75]。



HACK #61

6.1 比超人更聪明

闪电可以击中同一个地方两次，但可能性非常小。概率法则使我们能够计算一系列罕见事件连续发生的可能性。

我们偶尔会听到一些极不可能发生的事件多次发生在同一个人身上，比如一个在森林里漫步的人被雷电击中了7次，或是一对新泽西夫妇两次赢得了彩票大奖。当他们出现在新闻中时，这

些故事通常包含对当地统计教授的采访，教授会估计这种事情发生的几率。

计算一系列事件发生可能性的数学方法相当简单。难的是合理估计任意单一事件发生一次的概率。然后，你只需将这些单个概率相乘，就能得到全部怪异事件发生的可能性。

6.1.1 幸运的路易丝・莱恩

为了展示整个系列事件发生可能性的计算中涉及的步骤，我选择了文学名著中的例子。路易丝・莱恩（Lois Lane）漫画杂志的第56期描述了一系列罕见的事件，该期杂志由DC漫画在1965年4月出版。故事的通用模式是：路易丝有一些看似很难解释的超能力，但在故事的结尾会对此有某种简单的解释。



路易丝・莱恩，现在是漫画书中主人公超人的妻子（之前是他的女友和头号粉丝），是20世纪六七十年代DC漫画中非常受欢迎的人物。如今的资深漫画爱好者把那个时代的路易丝漫画作为独特奇怪的漫画创作来欣赏。路易丝几乎每天都能成功挑战概率。关于她的漫画应该被列为统计课程必读材料。

我们要讲的这个奇怪经历包含了一个统计Hack，在故事结尾，超人对其进行了解释。路易丝假装自己具有读心能力，借此待在绰号为“极小概率”的犯罪分子拉金身边，或许还能为她的报纸提供热点新闻素材。

一切进展顺利。她现在被拉金绑架了，拉金强迫她提供“读心能力”的信息，这样他就可以去犯罪。路易丝很幸运，犯罪分子也一样，她的盲目猜测是正确的，拉金让她一直活着。路易丝的猜测非常准确，以至于她自己都开始相信自己具有超能力。

最终，解救了路易丝的超人证明，路易丝只是幸运罢了！很幸运。令人吃惊的是，这是不可思议的幸运。路易丝正确预测冗长的系列和准确猜测的几率是极其微小的，她只是运气特好。恭喜，路易丝！

超人给出了路易丝实现这一梦幻般壮举的几率，但故事的作者（匿名）没有提供计算方法。让我们回顾一下路易丝做的随机猜测，我们自己计算，并检验一下超人的数学。为了确定这一系列的独立事件的概率，我们将应用乘法法则[Hack #25]。

6.1.2 猜测

在故事中，路易丝正确猜中了（注意，是完全随机地）如下问题：

(1) 5辆相同的装甲卡车，哪辆才是运送首都银行现金的；

(2) 公司里存放工资的保险箱密码；

(3) 未登记的、镇上首富的电话号码；

(4) 20 000棵树中，哪一棵的下面埋藏了银行抢劫犯的战利品。

超人救了她之后，她终于在猜测罐里软糖数量这道题上失败了。因为超人向莱恩女士解释说她没有超能力，他认为她随机猜中这4道题的几率是326 454 839 047（失败）比1（成功），或是1/326 454 839 048。

“我明白了，超人！”她说，“我很幸运地逮到‘那次机会’。”“是的，”超人说，“毕竟，也有人总是赢得彩票大奖（或某种类似的荒谬事情）。”由超人或他的超级电脑计算出的数字肯定是很大的，这好像是正确的，但我不认为这是正确的。我的猜测是，这个结果甚至会更加不可思议。

6.1.3　计算

让我们完成自己的计算。对于猜测1和猜测4，我们可以独立且非常准确地计算出猜中问题答案的几率。对于猜测2和猜测3，我们必须做一些假设。

下面是路易丝作出的猜测以及每个猜测几率的真实计算。



对于被要求给出一连串不可能事件的可能性解释的统计学家来说，这里所涉及的数学是工作中比较容易的一部分。困难的部分在于确定开始的值，以及确定等式的各个部分。正如你看到的，我们尝试估计路易丝有多么幸运，我们将要做一些适度狂野但合理的猜测，来或多或少估计任意特别事件发生的几率。大多数时候，统计学家无法准确获知事件的发生几率。他们往往专注于理论情况下事件发生的几率，而不关注莱恩女士那样的真实生活问题。

1. 猜测一

5辆相同的装甲卡车，哪辆才是运送首都银行现金的？这是最简单的一个问题。5种可能性，有1个是正确的选择。几率是1/5。

2. 猜测二

路易丝猜测存放大公司工资的保险箱的密码。这是一个真正的难题。路易丝不仅要猜中表盘需转至的5个数字，同时也得猜中5个不同数字必须遵循的顺序，并且要猜中转盘转动的方向。

在现实世界中，保险箱有多种不同类型的组合密码，所以很难确切知道我们应该对这个问题做什么假设。我对密码破译进行了一些研究（可以说是因为这个Hack），还对组合密码保险箱进行了一些了解。通常，密码箱共有1~8个数字的组合序列。我猜测，3个或5个数字的组合序列是最常见的。表盘上的数字可以是任意范围的值，但是0~99在大保险箱中很常见，比如故事中的工资保险箱。

所以，首先，比方说，她随机选择有3个或5个数字组合的保险箱。这个猜测的几率是1/2。假设她每次从0~99中随机挑选一个数字：序列中每个数被选中的几率都是1/100。她还必须猜测开始的方向。比方说，大多数保险箱（80%），开始是向左侧旋转的，而只有20%（1/5），开始是向右旋转（这是她的猜测）。

目前，一切良好。但是，路易丝实际的选择导致这一切非常棘手。她预测的是“11向右……13向左……5向左……向后旋到8……向前旋到15。”这是一个非常奇怪的组合。首先，我们通常会以另一种方式报出组合顺序：*向左*13，而不是13*向左*。第二，怎么可能连续两次向左旋转！毫无疑问，你必须改变表盘的方向以锁定序列中的每个数字。毕竟，表盘在其转动过程中会经过很多数字。它是如何知道是否要把每个经过的数字视作组合序列的一部分？我打算假装匿名作者误报了组合顺序了，否则，我不得不陷入无尽的混乱循环并因此被困此处，我的手指停在键盘上，再也无法继续下去。

最后，为什么路易丝不再说向左或向右，而开始说“后退”和“前进”？这只会使得她的方向不清楚。（难道是为了失败时为自己开脱？）再一次，我假定她使用这些术语意味着方向的变化，即使*后退*可能意味着*向左*，*前进*意味着*向右*，这也会使事情变得更复杂。那么，对于这样一个猜测的保守概率估计是1/2 × 1/5 × 1/100 × 1/100 × 1/100 × 1/100 × 1/100，即为1/100 000 000 000。

3. 猜测三

路易丝还猜中了未公开登记的、镇上首富的电话号码。有几种方法可用于计算。

首先，如果路易丝考虑得很简单（没有冒犯路易丝粉丝的意思，但我如此猜测），她可能只设置这样的限制因素：电话号码必须有7位数字，并且不以0开头。根据这些规则，有9 000 000个可能的电话号码。这意味着我们开始时假定有10 000 000个可能的7位数（9 999 999是最大的7位数，再加上一个数0 000 000）。

如果我们不将以0开头的数字算入其中，就消除了0 000 000这个数字和所有的6位数字或6位数以下的数字（有999 999个）。我们几乎消除了百万种可能。那么，在这种情况下，路易丝猜中数字的几率将是1/9 000 000。让我们给路易丝一些考虑时间，并假设她不会猜自己的电话号码或其他她记住的电话号码。我猜这样的号码可能有10个。那么，路易丝将从8 999 990中选择1个。

如果更聪明一些，路易丝（比方说，为了论述的必要）可能知道大城市正在使用的特定总机号、那些可能被用作未登记的号码或小城镇富裕人群的号码，等等。在以前，一个特定区域代码的前3个数字有可能是总机号。我们一般估计一个城市的人口规模有50万左右，所以她可能从这里面选择。在“更聪明的路易丝”场景下，她的胜算有了很大的提高。现在，她可能是从500 000个数字中盲目猜测，而不是从9 000 000个数字中猜测。她猜中的几率可能是1/500 000。我对路易

丝智商的粗略估计表明，这种情况不是最有可能的，但她是一名大城市报纸的记者，所以也许有这方面的知识。让我们仁慈一点，就这么假设吧。

4. 猜测4

最后，路易丝猜中了银行抢劫犯的战利品究竟埋在20 000棵树中的哪一棵下面。如同猜测一，这也非常容易计算。如果树林里有20 000棵树的下面真的可能埋有战利品（这个数字很可能是估计的或四舍五入的），那么正确猜测的几率是1/20 000。

6.1.4　最终概率

那么，假定路易丝没有错，知道保险箱和电话号码系统的各种事情，她在这4个问题上连续正确猜中的几率是1/5 × 1/100 000 000 000 × 1/500 000 × 1/20 000。保守地说，这个序列被幸运猜中的几率是1/5 000 000 000 000 000 000 000，甚至比现在已经令人难以相信的1/326 454 839 048还要引人注目。

“我明白了，超人！我很幸运地抓住了正确机会。”路易丝总结道。的确如此。当然，这个几率比将来超人向路易丝求婚的几率更糟糕，但那发生了。那么，我应该向谁诉苦呢，超人还是超人夫人？



6.2　揭秘惊人巧合

概率的模式会产生一些不寻常却有趣的一致性。下面教你如何解释那些看起来令人难以置信的巧合。

统计学家的职责偶尔也会令其伤感，其中之一就是把这个充满奇思妙想、美好意外发现以及不时冒出惊喜的世界，变成一个沉闷的、可预测的、无趣的地方。在这里，我也即将这么做，如果你宁可继续戴着乐观的眼镜，那现在就戴上它们，跳过这个Hack，选择另外一个Hack（我建议你选择更令人愉快的话题，比如赢得大富翁[Hack #51]）。

我选择科学性，并把世界视为理性的，建立在遵循因果链的结果之上。我的问题是（如果你和我具有相同的思考方式，也许你的也一样）：当我面对异常（很难解释的，意想不到的事情）时，很容易把异常发生作为某种神秘的、超自然的或某种意义上超出科学已知范围的证据。巧合是一个很好的例子。当我看到一个令人难以置信的巧合时，我忍不住就陷入非科学解释的舒适的坑里，如命运或共时性（synchronicity）。



共时性是开创性心理医生卡尔·荣格（Carl Jung）提出的术语，代表个人的有意义的巧合。他把它们视作对无意识内心世界的洞察，但不排除用伪神秘去解释它们。他不是一个统计学家。

我解决问题的方案是：思考一下，并应用概率的一些基本规则（也许这也是你的解决方案，如果你依然和我一样思考）。这样一来，我可以掌握真实情况，考虑到存在于宇宙中的大样本，把此类巧合视为是不可避免的。通过应用这些规则，我对我生活的世界感觉更好了。我可以在几率的怀里安然入睡，我不需要神秘的、神奇的解释。这里有3种策略可供你应对下一个遇到的惊人巧合。

6.2.1　比较可能结果的数量

当我还是个孩子时，我曾经在漫画书上看到一则广告（如，*Statboy和他的名叫Parameter的飞天狗*）。这则广告推销改变后的美国便士，便士不仅包括标准的林肯肖像，另外还有约翰·F. 肯尼迪的肖像。有一长串清单列出了这两位总统共有的"令人瞩目的"巧合，以便解释他们应该放在一起的原因（而且，我记得，如果我购买一整套便士，我甚至会得到一张小海报，海报上列出了这些相似点）。

该清单不仅包括显而易见的事实，比如两个人都被暗杀，继任者都是名为约翰逊的副总统，还包括了一些其他事实。我可以（的确）把这些巧合解释为两者之间某种重要的、有神奇联系的证据。让我们以这些巧合为例，将其作为一个研究问题：这两位总统之间是否存在不寻常的相似点？



*我现在突然想起来，当时那本漫画书上的广告促使我思考了一段时间，**巧合**（coincidence）一词源于单词**硬币**（coin）。当然，我很快就明白了（肯定是通过研究生院），这也只是一种巧合。*

当判断一个巧合是否令人瞩目或是否可预见时，有一个工具可供使用，那就是计算可能的结果数，然后判断给定的结果（巧合）是否会偶然发生。这是预测一群人中，是否有人同一天生日所用到的方法[Hack #45]。

表6-1的第一列介绍了在那些老漫画书广告以及一些"难以置信"的出版物中列出的一些巧合。第二列是一个简短的清单，列出这俩人可能相同、实际却不同的特征。

表6-1：比较亚伯拉罕·林肯和约翰·F. 肯尼迪

一些惊人的巧合	一些不起眼的非巧合
两人都被暗杀	身高不同
都在60岁当选	体重不同
刺杀肯尼迪的人从仓库中开枪，隐藏在剧院里；刺杀林肯的人在剧院里开枪，隐藏在仓库（嗯，至少是谷仓）	他们死时年龄不同（尽管他们出生时同龄）
林肯在福特剧院被刺杀；肯尼迪在福特车中被刺杀	他们的出生年份和日期不同
两人都在星期五被刺身亡	两人的中间名不同

（续）

一些惊人的巧合	一些不起眼的非巧合
两人都坐在妻子旁边被刺身亡	两人的妻子有不同的姓名，可能还有不同尺码的鞋子
两人的继任都叫约翰逊	两继任者的全名不同
	林肯留着胡子，肯尼迪没有（我想起来了，他们的脸型非常不同）
	肯尼迪偶尔可能会打保龄球，林肯一生从未打过一场保龄球比赛

如果只注意林肯和肯尼迪之间相对较少的一致性（命中），而忽略所有的、几乎无限多的不一致（非命中），就很容易误以为存在一些不可思议的连接。当然，依然可能存在一些不可思议的连接，但“巧合”无法为它提供证据。

6.2.2　找出实际几率

如果你遵守任意一种规则去玩扑克牌（如果你是一位名气不大的好莱坞名人，你显然一直在打牌），知道自己很少会看到皇家同花顺：同一花色的10、J、Q、K、A，共5张牌。如果你的对手被发到了一个皇家同花顺，那会引人注目吗？你会怀疑他作弊了吗？这取决于你一生中共看到多少扑克手牌，或你最近看到了多少手牌。

让我们用简单的5张牌做数学运算。为了计算发出的5张牌形成同花顺的几率，我们先计算出可能的5张手牌数，并和那些被定义为皇家同花顺的手牌组合数进行比较。这个过程需要3步。

(1) 考虑扑克牌的顺序，计算可能的手牌组合数。我们以这种方式开始是因为这样的数学计算最容易。52张牌中的任意一张都可能成为第一张发出的牌，然后剩余51张中的任意一张都可能是下一张发出的牌，再然后剩余50张手牌中的任意一张，以此类推，直到48张手牌中的任意一张。所以，当顺序有影响时，可能的手牌数是：

$$52 \times 51 \times 50 \times 49 \times 48 = 311\,875\,200$$

(2) 但是，顺序无关紧要。所以，我们用这个巨大的、所有可能的手牌总数除以可能的不同序列数。不同序列数是$5 \times 4 \times 3 \times 2 \times 1 = 120$，所以可能的5张扑克手牌数是：

$$311\,875\,200/120 = 2\,598\,960$$

(3) 因为只有4种花色，所以只有4种可能的皇家同花顺，我们用这个积极的结果（4）除以可能结果数（2 598 960），概率是0.000 001 539，即1/649 740。

每经历649 740次手牌，你的对手或你能被发到皇家同花顺的5张牌。所以，如果它确实发生了，那肯定是罕见的。如果它在同一场比赛中不止一次发生，你应该把它解释为非常惊人的巧合或是作弊的证据。你自己来决定是将其看为巧合还是作弊。我只知道我的计算和自己的猜测。



怎么能发到皇家同花顺？毕竟，在纸牌游戏和得州扑克中，玩家有机会改善他们的手牌或至少把它引向某个目标。在发牌中，如果你有4张牌能组成皇家同花顺，并希望放弃第五张牌而抽取一张新牌，你有1/47的成功几率，或0.021个百分点。如果你有两次机会来提高你的手牌，几率就上升至0.043%，即约每25次尝试成功1次。

6.2.3　移除分配给无意义事件的意义

当必须赋予数据意义时，人类的大脑处于最佳状态。我们令人瞩目的智慧甚至可以在没有意义的情况下找出意义。通常情况，我们会认为自己见证了一系列巧合奇迹。当我们寻找巧合时，我们就能看到巧合。

非常不可能的事件一直在发生：每一天、每一小时，甚至每一分钟。对于非常不可能的事件，只有当我们认为它们有趣时，它们才是有趣的。想想我们的扑克例子。因为有大约260万种可能的5张扑克手牌，任何特定手牌的几率约是1/2 600 000。我们认为有特别意义的手牌，比如同是黑桃的10、J、Q、K和A，和我们认为不具有特别意义的手牌，比如梅花4、黑桃6、方片J、黑桃K和红桃A，几率是一样的。为什么你对发到皇家同花顺和任何其他随机手牌组合的惊奇程度不同？对所有扑克手牌来说，概率是相同的。我们对特定的结果赋予了意义。

下一次，你在一个拥挤的地方（如，棒球比赛场、游乐园或机场）遇到某个认识的人，你认为这种巧合有意义只是因为你碰巧认识这个人。是的，你会遇到某个特定的人（除非你被跟踪）的几率非常渺茫，但你100%会遇到其他人。所有其他人只是碰巧与你在相同时间出现在相同地点。它是一个巧合，个体的这种特定组合在同一个时间同一地点发生是非常不可能的。但是，这对你来说不是一个有意义的巧合。



如果我们算上你认识的每一人，那么你遇到熟人的几率会更高。比方说，你认识200人，你某天晚上自己去堪萨斯城皇家队看棒球比赛。如果那200个人每人每赛季去看一次皇家队比赛，每个赛季有81场主场比赛，200人中每个人有1/81的几率和你在同一个晚上出现在那里。那个时候你不太可能会遇到特定的人，比如你的叔叔弗兰克，但那里非常可能有你认识的人。大约有92%的几率，你200个好友中的一个或多个会在那里，即使他们每个人都很少去看比赛。即使你只认识56个人，这56人中的一个或多个出现在那里的几率也是大于50%的。

我们每天都经历大量事件，人和事以非常不可能的方式交互和巧合。有时，这些巧合对我们有意义，所以我们注意到它们。但令人诧异的是，我们没有花费更多的时间关注这些非常不可能的事件。



6.3　识别生活中真正的随机

在你指责赌场经营不正当的赌博，或威胁你的老板你将起诉他只雇佣白肤金发碧眼的女人之前，这里有一个工具，可用来分离那些看起来非随机但可能随机的情境以及那些看起来非随机但可能没有随机发生的情境。也许吧。

随着你越来越深刻地意识到几率在你周围世界扮演着重要角色，你开始习惯性地对每天的情境进行统计分析，并可能对看起来不正确的模式过于敏感。但是，不要滥用你新发现的力量，把概率视为确定性。此外，不要错误地期望人们认为随机的事件看起来也是随机的。

6.3.1　随机是怎样的

看起来随机和真正的随机是不一样的。当事件有不同的可能结果，而每个结果有等同的发生几率时，其中任何一个都有可能发生。但是，人们的一般思维是，有若干同几率结果的事件，其最终结果应该看起来是某种方式，在一定程度上，这种方式看起来也是随机的（不管那意味着什么）。

举例来说，现实世界的研究发现，人们往往认为翻转硬币时，最可能的结果是那些看起来最为混杂的结果。为了说明这个观念，请看表6-2。（在没进行深入阅读前，不要看表6-3）你认为哪个确切的顺序最有可能发生？

表6-2：硬币翻转模式（不显示概率）

答　案	正面和反面的模式	概　率
A	正面、反面、正面、正面、反面	?
B	反面、反面、反面、反面、反面	?
C	正面、正面、反面、反面、反面	?
D	正面、正面、正面、正面、反面	?

很多人给出的答案是“A”。也许你给出的也是这个答案。当被要求解释为什么A看起来是最可能出现的结果时，可能有以下这样的解释。

- “其他的都太有顺序了。”
- “A更混杂，所以它的可能性比较大。”
- “A看起来更随机，就像它可能真的会发生一样。”

即使你知道抛硬币是随机的（假设硬币没有被加重），看起来随机并不使得某事更有可能。所有这些抛硬币的模式实际上具有同样的可能性，如表6-3中的数学所示。

表6-3：抛硬币模式（显示概率）

答案	正面和反面的模式	概　　率
A	正面、反面、正面、正面、反面	1/2 × 1/2 × 1/2 × 1/2 × 1/2=1/32=0.031 25
B	反面、反面、反面、反面、反面	1/2 × 1/2 × 1/2 × 1/2 × 1/2=1/32=0.031 25
C	正面、正面、反面、反面、反面	1/2 × 1/2 × 1/2 × 1/2 × 1/2=1/32=0.031 25
D	正面、正面、正面、正面、反面	1/2 × 1/2 × 1/2 × 1/2 × 1/2=1/32=0.031 25

当被要求预测抛一系列硬币的特定结果时，所有可能的结果一定是相同概率的，因为每次抛硬币都是相互独立的。换言之，硬币不知道它上一次是头着地还是尾着地，硬币也没有办法知道它下一次被抛出时哪一面着地。一枚硬币，像骰子或轮盘赌一样，没有记忆。

6.3.2　如何识别随机结果

当你看到不同寻常的事件时，想要知道它是否为不寻常的事件，你需要确定你关注的是组合还是排列。在概率论中，我们讨论概率的计算时要分清是某种组合的概率（例如，以任何顺序出现的3个正面和2个反面），还是某种排列的概率（会产生3个正面和2个反面的确切序列，如正面、反面、正面、正面、反面，以这个特定顺序出现）。

如果你被问到，哪个结果是最有可能的，或一个给定结果是否可能偶然发生，首先要确定你被问的是可能的组合（例如，以任何顺序出现正面和反面的总数，或是以不同方式发到相同花色的5张牌的总数）还是可能的排列。下面是两者的重要区别。

- 组合

组合是指当从某个总体中随机抽取时，能使结果达到某个特定数值的总方法数。硬币翻转就是从由50%正面和50%反面构成的、理论上无限大的总体中抽取的样本。组合的数量会有变化，这取决于感兴趣的特定值的数量。换句话说，对于抽5张牌或翻转硬币，抽到3张人头牌的方法比抽到5张人头牌方法要多。因此，抽到3张人头牌比抽到5张人头牌更有可能。

- 排列

排列是指给定数量的元素能以多少种方式排列。换句话说，它们是精确的序列数。在我们的硬币翻转例子中，5个元素，每个元素都有2种可能，进而导致32种不同的可能顺序结果。所以，表6-3所示的每个排列会发生的几率是1/32。

6.3.3　如何计算组合

可能的组合数量是通过把抽取的可能值的数目（例如，一枚硬币有2个值：正面或反面），在每次抽取时和它自身相乘得出：

值的数量^{抽取次数}

5次硬币翻转，有32种可能的组合（2^5）。

从总体中抽取特定元素的特定值（如3个正面）的方法数，计算方程如下：

$$\frac{n!}{r!(n-r)!}$$

这个方程，需要这些变量：

n 元素或抽取的数量（例如，5次硬币翻转）。

r 感兴趣的特定抽取（例如，3个正面）。

! 阶乘，表示这个数乘以比此数小1的数，然后乘以比此数小2的数，依此类推，直到最后乘以1。例如，5！代表5×4×3×2×1=120（顺便说一下，这就是在扑克手牌中，为什么5张牌有120种可能组合[Hack #62]）。

那么，5次硬币翻转获得3个正面的方式数为：

$$\frac{5!}{3!(5-3)!}=\frac{120}{6(2!)}=\frac{120}{12}=10$$

32种可能的组合中选出10种组合，意味着你通过5次抛硬币正好得到3个正面的几率是10/32，或约31%。

在一个荒岛上进行统计黑客

如果你在一个荒岛上，没有书籍或方程方法，但必须找出5次硬币翻转中正好出现3次正面的频率是多少时，你可以使用粗略近似的方法：把所有可能的翻转模式列出来，并数出它们之中有多少正好有3个正面。它会如下面这样，符合要求的结果（3次正面）以粗体显示：

HHHHH THHHH HHHHT **THHHT** **HHTTH**THTTH HHTTT THTTT HHHTH **THHTH** **HHHTT** THHTT HHTHH **THTHH** **HHTHT** THTHT HTHHH **TTHHH** **HTHHT** TTHHT HTTTH TTTTH HTHTT TTHTT **HTTHH** TTTHH HTTTT TTTTT **HTHTH** TTHTH HTTHT TTTHT

6.3.4 什么时候需持怀疑态度

判断一个模式是否随机（即，什么被期望为偶然出现的），需要：

- 知道某种组合的几率（不是排列）；
- 克服“期望随机结果不会产生可识别模式”的心理预期；
- 在质疑数据前，设置事件需达到的不可能发生的概率标准。

让我们回到硬币翻转表，现在表6-4新增了感兴趣的结果概率。

表6-4：硬币翻转结果和概率

顺序	顺序概率	结果	结果概率
正面、反面、正面、正面、反面	0.031 25	3个正面	0.312 50
反面、反面、反面、反面、反面	0.031 25	5个反面	0.031 25
正面、正面、反面、反面、反面	0.031 25	3个反面	0.312 50
正面、正面、正面、反面、正面	0.031 25	4个正面	0.156 25

这些结果中，最稀有的是5个正面，5枚硬币翻转，100次中会有3次出现5个正面。在一个给定的尝试下，这不太可能发生，但在一系列尝试中，它偶尔会发生。如果在一系列的尝试中经常发生，其中一定有某种原因。

你习惯什么水平的可能性？事件得有多罕见，你才能判断其不是偶然发生的？科学家们已设定了5%的标准。如果研究表明，这个结果偶然出现的几率只有5%或更少，那它通常被认为是显著，可能作为有几率以外的因素在起作用的证据。

不过，当你想指责某人是骗子时，你必须自己决定。祝你做决定时好运！它导致打架的几率应该小于5%。

——吉尔·罗米尔和布鲁斯·弗雷



6.4 识别伪造数据

如果你之前没有对数字进行太多思考，你可能很自然地假设在最随机的数据集中，所有数字出现的可能性等同。但根据本福特定律，对于许多类型的自然发生的数据，数字越小，它以首位数出现的频率越高。你可以用这个秘密知识来检验任何数据集的真实性。

在电子计算器时代远未到来的19世纪，科学家利用出版书籍里的表格发现了对数的值。一位特别细心的19世纪天文学家、数学家西蒙·纽科姆（Simon Newcomb）发现，含有对数表格的页面，其前几页比后几页更加破旧。纽科姆下结论，以1开头的数字出现的频率比以2开头的数字更高，以2开头的数字出现的频率比以3开头的数字更高，以此类推。

根据他的观测，纽科姆于1981年在《美国数学杂志》（*American Journal of Mathematics*）上发表了实证结果，其中阐述了许多类型的自然产生数据的概率，数据以d开头，d=1,2,⋯,9。纽科姆的首位有效数法则（first significant digit law）几乎不被关注，在很大程度上甚至被遗忘了，直到50年后，就职于通用电气公司的物理学家富兰克·本福特（Frank Benford），注意到了同样的对数表破损模式。

经过对大量数据的广泛测试（20 229个观测结果）——包括原子量、河流的排水区、人口普查数字、棒球统计数据和财务数据，本福特将首位有效数字概率定律发表在美国哲学会的会议记

录上（本福特，1938）。这一次，首位有效数字法则吸引了更多的关注，被称为本福特定律。尽管文章于1938年发表后，本福特定律变得众所周知，且其中包括大量的统计证据，但它缺乏一个严谨的数学基础，直到1996年，佐治亚理工学院的数学教授西奥多・希尔（Theodore Hill）才提供了严谨的数学证明（希尔，1996）。

今天，本福特定律在多个自然产生数据的领域中有着常规的应用。也许本福特定律最实际的应用是检测会计中的欺诈数据（或无意的错误），由圣迈克尔学院（Saint Michael's College）工商管理和会计系的教授马克・内格罗尼（Mark Nigrini）率先应用（http://www.nigrini.com/）。

伪造数据的检测不仅在会计中非常重要，而且在各种各样的其他应用中也很重要（例如，在药物测试的临床试验中）。本Hack介绍了本福特定律，告诉你如何应用它，提供了一些直观理由证明其有效性，并给出说明什么时候能运用本福特定律的指导原则。

6.4.1　如何生效

在最简单的形式中，本福特定律指出，在许多自然产生的数字型数据中，第一个（非零）有效数字的分布遵循对数概率分布，如下所示。沿用希尔的用法（1997），令 $D_1(x)$ 表示十进制数字 x的首位有效数字。例如，$D_1(9108)=9$，$D_1(0.025708)=2$。

然后，根据本福特定律，$D_1(x)=d$ 的概率可按下式计算，其中d可以等于1,2,3,⋯,9：

$$P(D_1=d)=\log_{10}\left(1+\frac{1}{d}\right)$$

这样，表6-5给出了首位有效数字的概率。

表6-5：本福特定律下的首位有效数字的概率

第一个非零数字	本福特定律下的概率
1	0.301
2	0.176
3	0.125
4	0.097
5	0.079
6	0.067
7	0.058
8	0.051
9	0.046

6.4.2　验证定律

为了证明本福特定律，我会考虑两个你可以自行验证的例子。

1. 街道地址

付诸行动验证本福特定律，打开你所在城市或城镇的电话簿，翻到任何一页，记录下以非零开头的每个十进制数门牌号码。两页就足够了。除非你所在的城镇有些不寻常，不然相对频率应和通过本福特定律预测的概率相似。

表6-6显示了413个家庭门牌号码的计算结果，号码取自2005年至2006年Narragansett/Newport/Westerly这一地区的RI黄皮书（白页部分）中的两页。

表6-6：遵循本福特定律的地址

首个非零数字	门牌号首个数字的相对频率	基于本福特定律的概率
1	0.334	0.301
2	0.174	0.176
3	0.143	0.125
4	0.075	0.097
5	0.073	0.079
6	0.075	0.067
7	0.046	0.058
8	0.043	0.051
9	0.036	0.046

图6-1更清楚地展示了这个模式。

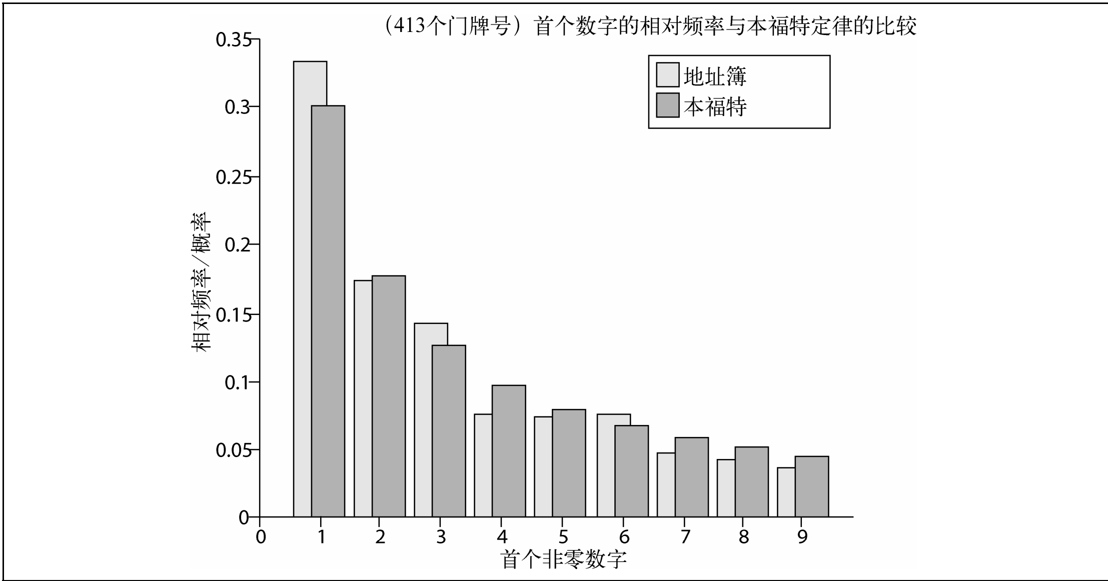


图6-1：遵循本福特定律的街道地址

虽然实际情况和本福特定律不完全一致，但你可以看到一个合理的良好匹配。如果你采用更

大的地址样本，由此产生的相对频率会更接近本福特定律预测的频率。

2. 股票价格

股市遵循本福特定律。你可以在http://quotes.nasdaq.com/reference/comlookup.stm上获取最新的纳斯达克证券价格，进而自行验证。

图6-2和表6-7显示了2006年1月27日纳斯达克证券第一个非零十进制数字的相对频率，并和本福特定律所预测的概率进行对比。

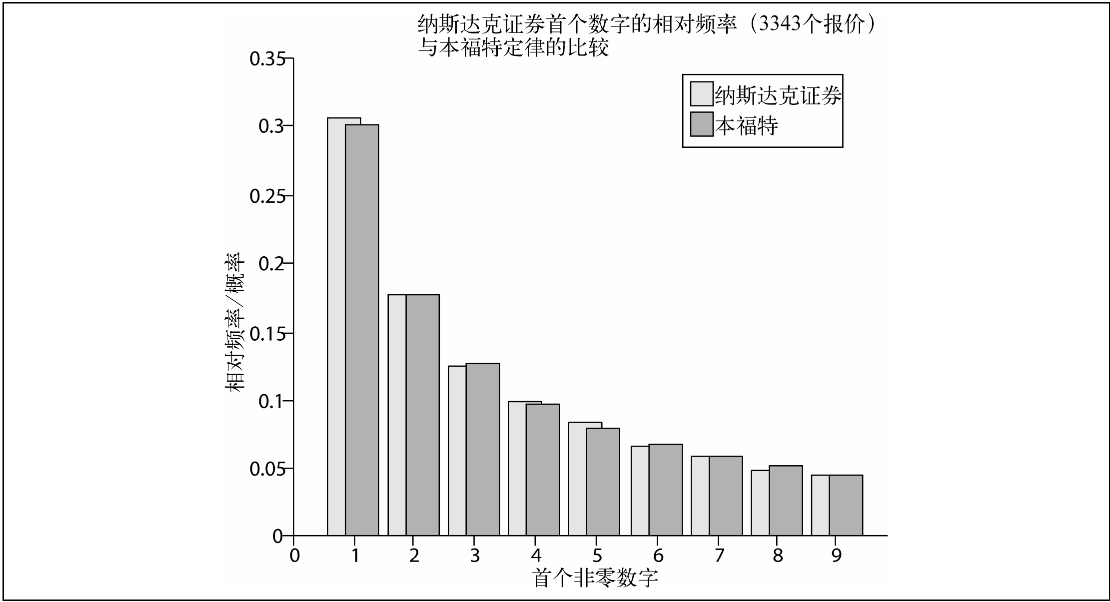


图6-2：遵循本福特定律的股市

表6-7：遵循本福特定律的纳斯达克证券

首个非零数字	纳斯达克证券首个数字的相对频率	根据本福特定律的概率
1	0.301	0.301
2	0.167	0.176
3	0.133	0.125
4	0.095	0.097
5	0.082	0.079
6	0.071	0.067
7	0.055	0.058
8	0.045	0.051
9	0.049	0.046



你可以在http://homepage.mac.com/samchops/benford/上得到用于生成这部分表格和数字的Matlab代码。此外，Mark Nigrini在http://www.nigrini.com/datas_software.htm上提供了DATAS软件（包括一个免费的学生EXCEL程序），可执行对第一个、第二个或前两个数字的更复杂的数据分析。

6.4.3　本福特定律更普遍的应用

本福特定律并不只适用于首个非零数字，也同样适用于其他数字的概率。再次，遵循我们之前讨论的方式，令 $D_2(x)$ 表示十进制数字x的第二个有效数。例如，$D_2(9108)=1$ 、$D_2(9018)=0$ ，而 $D_2(0.025108)=5$ 。

注意，不同于首个有效数字，第二个有效数字可以是零。

然后，根据本福特定律，$D_2(x)=d$ 的概率，由下面的等式给出，其中d可以等于0,1,2,⋯,9：

$$P(D_2=d)=\log_{10}\left[\left(1+\left(\sum_{i=1}^{k}d_i\times 10^{k-1}\right)^{-1}\right)\right]$$

这个公式得出了第二个有效数字的概率，如表6-8中所示。

表6-8：本福特第二数字定律

第二个有效数字	根据本福特定律的概率
0	0.119 68
1	0.113 89
2	0.108 82
3	0.104 33
4	0.100 31
5	0.096 68
6	0.093 37
7	0.090 35
8	0.087 57
9	0.085 00

从表6-8可以看出，不同于相应的首位数字，第二个有效数字之间的概率差异不那么具有戏剧性。

现在，回到股市。为了论证本福特定律和第二个有效数字有关，我计算了之前纳斯达克证券的第二个有效数字的相对频率。结果如表6-9所示，再一次说明确实与本福特定律有密切的一致性。

表6-9：遵循本福特第二数字定律的纳斯达克证券

第二个数字	第二个数字的相对频率	根据本福特定律的概率
0	0.128 03	0.119 68
1	0.114 27	0.113 89
2	0.109 18	0.108 82
3	0.102 90	0.104 33
4	0.102 30	0.100 31
5	0.092 73	0.096 68
6	0.090 64	0.093 37
7	0.091 53	0.090 35
8	0.084 06	0.090 35
9	0.084 36	0.085 00

本福特的一个更普遍的概率公式可以用来计算第n位的相应概率。设$D_k(x)$表示十进制数字x的第k个有效数字。然后，根据本福特定律，$D_1(x)=d_1$，$D_2(x)=d_2$，…，和$D_n(x)=d_n$的概率由下面的公式给出：

$$P(D_1=d_1,D_2=d_2,\cdots,D_n=d_n)=\log_{10}\left[\left(1+\left(\sum_{i=1}^{n}d_i\times 10^{n-1}\right)^{-1}\right)\right]$$

注意，如果k不等于1，那么d_k可以等于0,1,2,…,9，正如前面所指出的，d_1可以等于1,2,…, 9。

6.4.4　其他生效领域

本福特定律的两个独特性质是尺度不变性和底数不变性。

1. 尺度不变性

本福特定律的尺度不变性是指，如果你用某个数乘以任何非零的常数，你依然会得到接近于遵循本福特定律的分布。因此，你以美元、第纳尔[1]或谢克尔[2]衡量股票的报价，以英里或公里测量河流的长度，都没有区别。你最后总是会得到遵循本福特定律的数据。

为了证明这一点，我使用前面例子中的纳斯达克证券数据，把每个值都乘以p。正如你在表6-10中看到的那样，相对频率仍遵循本福特定律。

注1：第纳尔是南斯拉夫、伊拉克及阿尔及利亚等国的货币单位。——译者注

注2：古希伯来或巴比伦的度量单位和钱币。——译者注

表6-10：遵循本福特定律的扩大后的纳斯达克证券

首个非零数字	纳斯达克证券首个数字的相对频率	根据本福特定律的概率
1	0.306	0.301
2	0.176	0.176
3	0.123	0.125
4	0.097	0.097
5	0.081	0.079
6	0.066	0.067
7	0.058	0.058
8	0.049	0.051
9	0.045	0.046

2. 底数不变性

本福特定律的底数不变性是指，它不仅适用于底数10，而且还适用于更一般的底数。此外，西奥多·希尔发现，本福特定律是唯一具有这一性质的概率定律（希尔，1995）。



你可以在希尔（1997）论著里找到一般底数的本福特定律公式。6.4.7节有关于此出版物的详细说明。

数据具有以下特点时，本福特定律的效果最佳。

- 足够的可变性

该变异越高，本福特定律的运用效果越好。

- 无内置最大值或其他类似的约束

例如，本福特定律并不适用于高年级学生的年龄，或当地老年人中心的成员。

- 数字来自于计数或测量

例如，它不适用于社会安全号码和邮政编码，因为它们是简单的识别码，不是真正的数值。

- 大样本

数据集越大，本福特定律的运用效果越好。

- 随机抽样

数据来自于大量的、随机选中的、符合概率分布的随机样本。随机抽样的实现为希尔证明本福特定律提供了有力的支撑（贝克尔，2000；希尔，1999）。

由于税收数据很好地遵循本福特定律，所以这一定律已经十分成功地用于识别虚假的纳税申报。在描述本福特定律的一些基本特征时，我们展示了如何对数据的违规行为进行迅速且随意的

检验。具体来说，任何人都可以很容易地计算第一个数字的相对频率，把这个结果和由本福特定律预测的结果放在一起，并进行仔细的对比检查。

在实际应用中，专家和权威人士使用的、用来确认偏离本福特定律结果以及其他违规行为的程序是相当复杂的。与本福特定律存在偏差并不能证明存在欺诈行为，但它确实给出了显著性，提示需要进一步调查，记住这点也同样重要。



你可参看内格罗尼（1996）了解更多利用本福特定律发现舞弊的细节，其中包括“拟合优度”测试。6.4.7节有出版物的具体信息。

6.4.5　生效原理

尽管对本福特定律的证明是相当具有技术性的，但也有一些针对此数学原理的、有见地且直观的解释。马克·内格罗尼（1999）就提供了一个这样的解释，我觉得特别有吸引力。

他的解释是这样的。想象一下，将初始金100美元用于某种投资，预计金额以10%的年增长率增长，总金额的第一位有效数字变为2将大约需要7.3年的时间。这是因为总量需要增加100%，才能达到200美元的价值。相比之下，考虑500美元增加到600美元的时间。如果我们继续假设10%的年增长速度，它将需要大约1.9年才能达到600美元。所以，投资金额的首位数是5时，所需时间比投资金额首位数是1时要少很多。一旦总金额达到1000美元，在其第一位数变成2之前（另一个100%增长），将再次需要约7.3年的时间。

现实世界确实比较复杂一点，但是这确实有助于解释为什么1作为首位数比其他数字作为首位数要更常见。另一种直观的解释是，如果有比大城市数量更多的小城镇，那么就有比长河流数量更多的短河流。

6.4.6　无效领域

本福特定律不太可能运用在无足够变异的数据集，或非随机选择的数据集上。例如，计算机上的文件大小大致遵循本福特定律，但只有当所选文件的类型没有限制时，才可以采用本福特定律。

为了说明这一点，我在一台苹果PowerBook G4笔记本电脑中找出了文件大小的第一个数字的频率。图6-3和表6-11中展出的结果显示出了本福特定律。

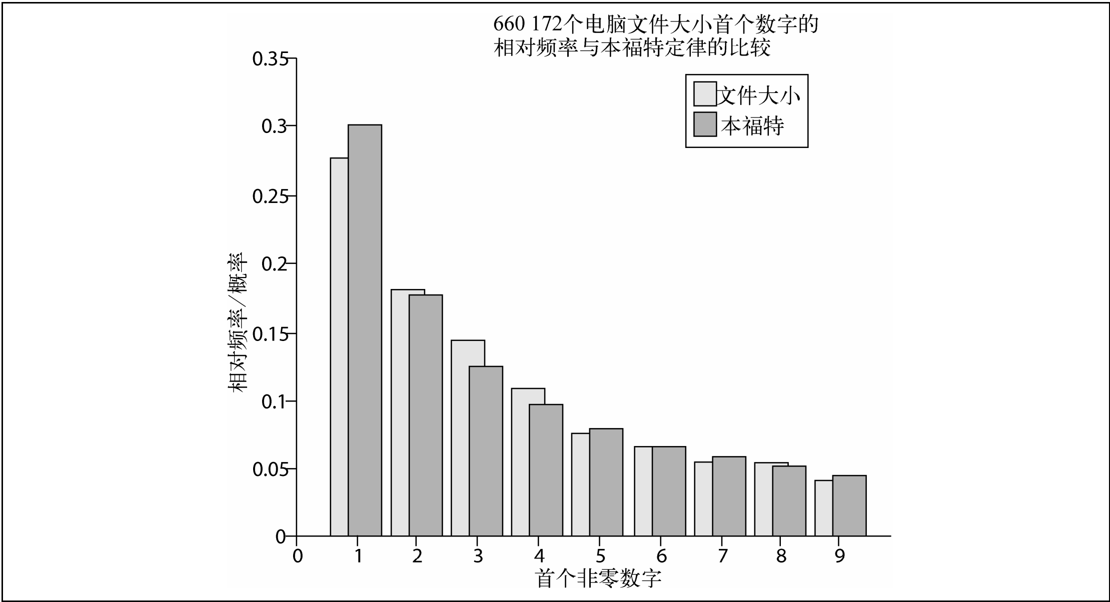


图6-3：遵循本福特定律的电脑文件

表6-11：大致遵循本福特定律的电脑文件

首位非零数字	660 172个电脑文件首位数字的相对频率	根据本福特定律的概率
1	0.277	0.301
2	0.181	0.176
3	0.144	0.125
4	0.107	0.097
5	0.076	0.079
6	0.067	0.067
7	0.054	0.058
8	0.054	0.051
9	0.041	0.046

尽管图6-3和表6-11所示的结果是基于660 172个文件的，表6-12显示，600个样本就足以表现出本福特定律模式（虽然不如更大样本的表现效果），只要文件样本是随机的。

表6-12：600个计算机文件大小的随机选择

首个非零数字	600个电脑文件首位数字的相对频率	根据本福特定律的概率
1	0.262	0.301
2	0.187	0.176
3	0.147	0.125
4	0.107	0.097
5	0.069	0.079

（续）

首个非零数字	600个电脑文件首位数字的相对频率	根据本福特定律的概率
6	0.070	0.067
7	0.052	0.058
8	0.057	0.051
9	0.052	0.046

为了便于比较，我计算了同一台计算机上iTunes音乐库中的MP3文件的相对频率。表6-13和图6-4表明，该组文件不遵循本福特定律。

表6-13：不遵守本福特定律的MP3音乐文件

首位非零数字	601首MP3文件首位数字的相对频率	根据本福特定律的概率
1	0.080	0.301
2	0.097	0.176
3	0.276	0.125
4	0.270	0.097
5	0.161	0.079
6	0.070	0.067
7	0.023	0.058
8	0.013	0.051
9	0.001	0.046

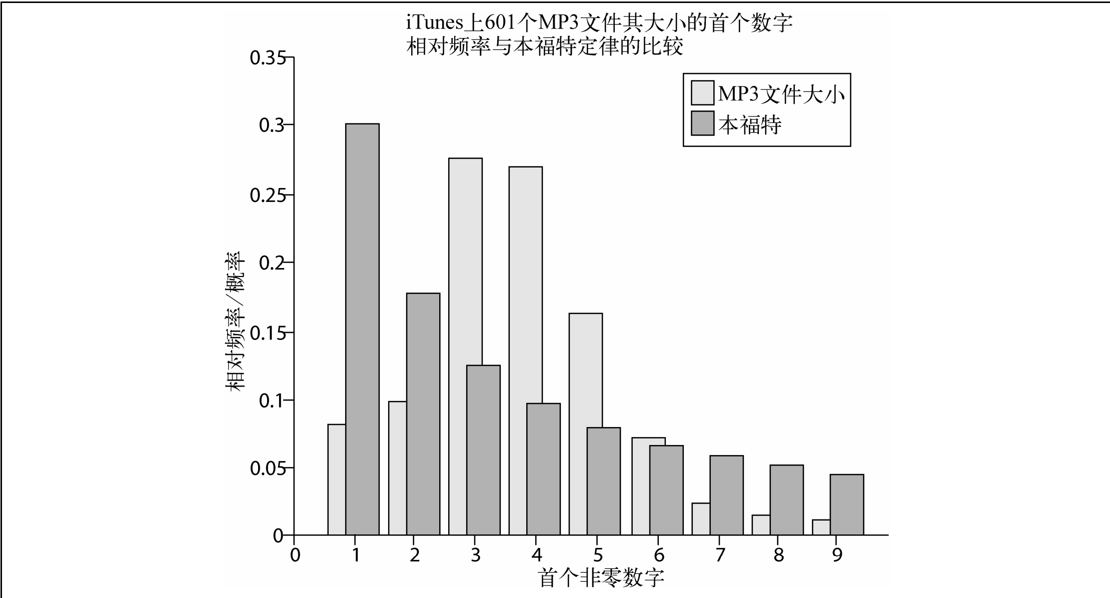


图6-4：不遵守本福特定律的MP3音乐文件

600首MP3格式的音乐文件的大小不近似本福特定律，这并不奇怪，因为MP3音乐文件的大小表现出的变异比更随机任取的600个计算机文件的变异要少得多。

6.4.7　参阅

- Becker, T. J. (2000). “Sorry, wrong number: Century-old math rule ferrets out modern-day digital deception,” *Georgia Tech Research Horizons*, http://gtresearchnews.gatech.edu/reshor/rh-f00/math.html.
- Browne, M. (1998). “Following Benford’s law, or looking out for no. 1.” *The New York Times*, August 4, 1998.
- Fawcett, W. (n.d.). “Significant figure generator.” http://williamfawcett.com/flash/SigFigDistbGen.htm.
- Benford, F. (1938). “The law of anomalous numbers.” *Proceedings of the American Philosophical Society*, 78, 551-572.
- Hill, T. P. (1996). “A statistical derivation of the significant digit law.” *Statistical Science*, 10, 354-363.
- Hill, T. P. (1995). “Base-invariance implies Benford’s law.” *Proceedings of the American Mathematical Society*, 123, 887-895.
- Hill, T. P. (1997). “Benford’s law.” *Encyclopedia of Mathematics Supplement*, 1, 112. Kluwer.
- Hill, T. P. (1999). “The difficulty of faking data.” *Chance*, 26, 8-13.
- Newcomb, S. (1881). “Note on the frequency of use of the different digits in natural numbers.” *American Journal of Mathematics*, 4, 72-40.
- Nigrini, M. (1999). “I’ve got your number: How a mathematical phenomenon can help CPAs uncover fraud and other irregularities.” *AICPA Journal of Accountancy Online Journal*, May 1999,http://www.aicpa.org/pubs/jofa/may1999/nigrini.htm.
- Nigrini, M. (1996). “A taxpayer compliance application of Benford’s law.” *Journal of the American Taxation Association*, 18, 72-91.

- 你可以在http://homepage.mac.com/samchops/benford/获得生成本部分图表的Matlab代码；在http://www.mathworks.com下载运行代码的Matlab安装包。

——欧内斯特·罗斯曼



HACK #65

6.5　物归其主

文体测算（Stylometrics）作为一种统计方法，可标识出定义作者风格的相关维度。它采用因素分析的方法来判断谁是作品的作者。

6

豪–马奇教授面临着一个问题。他最好的两名学生现在都坐在他的办公室里，希望能解决一个争议。豪–马奇博士将保罗的期末论文评为A+（这是一篇探讨巧克力牛奶重要性的历史论文），但问题是，丽莎声称那篇论文是她写的。这构成了抄袭指控！两人都是好学生，在过去都为教授写了许多高质量的论文。所以，判断谁是真正的作者并不容易，意识到最喜欢的学生之一是个骗子也不容易。

幸好，相比他担任的州立社区学院和货运学校兼职教授一职，作为优秀哲学博士的多年经验使他能够想出更有效的方法。除了一些不明显的统计爱好，豪–马奇博士还涉足*文体测算*领域，这是一种对文字作品风格分类的统计方法。该方法也可用于识别匿名作者。当有好几种可能性或者若干嫌疑人以供选择时，当嫌疑人的典型写作风格已知并已量化时，该方法的效果最好。让我们看着心碎的教授如何应用这些技术找到真正的作者。

6.5.1　建立模型

首先，豪–马奇博士让保罗和丽莎带来所有他们在过去写的、没有争议的其他论文。仅用短短几分钟，这些论文就被扫描到计算机中，并形成两位作者使用的不同单词的数据库。



或者，可以将论文以电子版的形式发送给教授，这样就无需扫描了；这和故事没有一点关系，那你为什么要问我呢？

第一步分析，将两位作者写的所有单词放在一起。豪–马奇博士数出每个单词的使用频率，在单词结合数据库中确定最常使用的50~100个单词。这些单词作为项目或关键变量构成*因素分析*（factor analysis）所用的数据。因素分析是这样一个统计方法：它着眼于组间变量的相关性[Hack #11]，并识别出一组群变量，这组群变量的彼此相关性比它们和其他变量的相关性更强。不管这些变量的共同之处是什么，它们都被假定共享一个因素、部分或维度。

便于我们故事的开展，我只列出10个豪–马奇博士认定的两位作者最常用的词。表6-14显示了这些词和它们的使用频率。当查看保罗和丽莎写的所有词时，“the”的使用频率为4.2%，“weasel”的使用频率为1%，以此类推。

表6-14：保罗和丽莎的常用词及其频率

词	频率
the	4.2%
and	2.1%
to	1.8%
a或an	1.2%
weasel	1.0%
of	0.8%

（续）

词	频率
in	0.8%
that	0.5%
it	0.4%
not	0.2%

这些词作为变量试图找出描述一个或多个风格维度的潜在因素。保罗和丽莎的风格可能体现在这些维度的不同地方。可能只有一个维度或因素导致这些词用法各异，也可能有很多维度或因素。一旦确定这些由相关变量共同定义的维度或维度上的载荷，任何写作样本都可以被放置在由因素搭建出框架的理论空间里。

豪–马奇博士进行因素分析的数据来自作品样本的各部分，每部分包含500个单词。每部分在每个单词变量上都有一个得分。得分是这个单词在该段落使用的次数。表6-15展示了豪–马奇博士收集的数据例子。

表6-15：研究数据的样本

	the	and	to	a/an	weasel	of	in	that	it	not
第1部分	21	8	11	5	4	0	0	1	0	2
第2部分	10	7	15	5	2	10	1	0	0	0
第3部分	5	5	5	2	6	12	2	4	1	0
第4部分	0	2	4	3	1	4	6	8	1	0
第5部分	4	11	16	2	0	3	5	0	3	1



表6-15中，分数表示每个单词出现在文本部分的次数。

6.5.2 因素分析

接着，豪–马奇博士进行因素分析，因素分析是一个相当复杂的数学过程，所以现在基本使用计算机来完成，与此同时，研究人员根据相关理论在分析过程中的不同时刻作出决策。基本上，要不断分析变量之间的关系，直到发现少数变量组似乎可以尽可能多地解释数据的变异性时，因素才被确定下来。每个分组变量共享的共性提供了定义该因素的数学素材。一旦因素被选择，任何观测（在本例中是文本样本）都能得到因素得分，然后以因素分数为坐标，将其置于那个理论空间里。

在本例中，分析表明，有两个因素很好地描述了样本文本。因素1通过使用的单词来定义，比如一端使用“a/an”而另一端使用“of”和“in”。换句话说，文本部分基于他们使用冠词的频次而不同，有较高冠词使用频率的部分往往使用较少的介词。因素2通过“weasel”一词的使用

频率来定义。

在探索性因素分析中，通常研究者对发现和命名能解释人类行为和特征的基本结构（即无形的特征）感兴趣。不过，在本例中，豪–马奇教授只对定义维度（例如，单词使用）感兴趣，这些维度是基于变量的，且能在两端对变量进行锚定。他没兴趣搞清楚为什么那些经常出现单词“the”的文本也也往包含高频率的“a”或“an”。他同样对“weasel”一词的使用为什么能够区分不同的写作样本不感兴趣。对他而言，他只需要知道这两个因素提供了一对良好的坐标轴，定位出两位作者在他们样本中使用的所有单词的位置。

计算保罗和丽莎样本论文的因素得分，很明显，结果表明两位作家有不同的风格。丽莎比保罗更频繁地使用“weasel”这个词，她的论文在因素2上得分高。丽莎的论文也倾向于高频使用冠词，在因素1上的分数也非常高。另一方面，保罗的论文往往避免使用“weasel”这个词，而且倾向使用因素1末端的介词。

仅使用单词来描述或许很难把握，所以我们借助一个图例画一幅图来演示样品文本的位置。图6-5显示了这两个因素：定义它们的单词使用，还有不同写作样本载荷在两个因素的位置。为便于讨论，图6-5只显示了少数的写作样本，只标出了表6-14和表6-15中的10个词。图中同样标出了那篇有争议的论文在理论空间的维度位置。

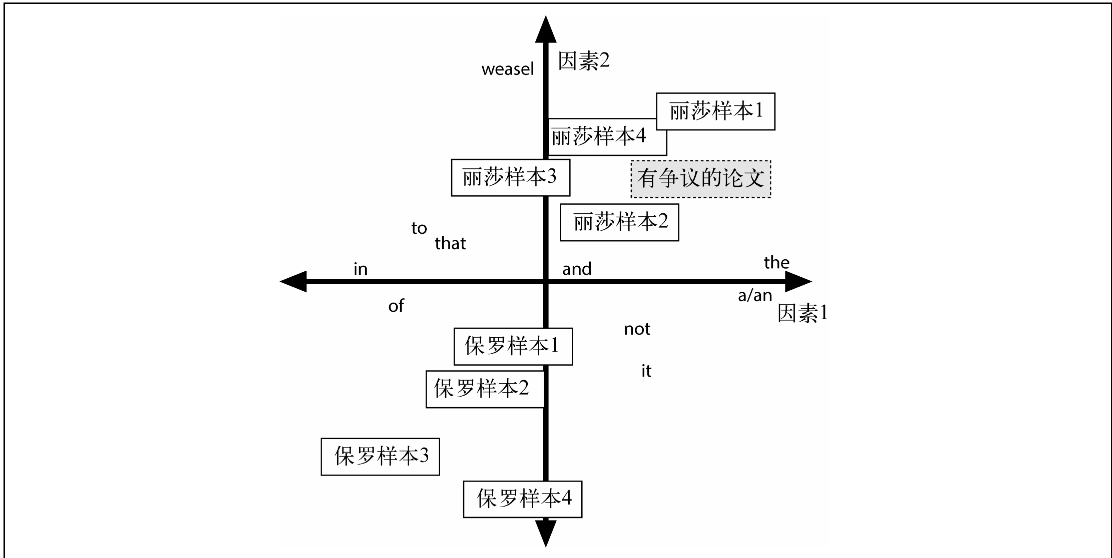


图6-5：文本样本的因素分析

谜题的答案现在已经很清楚了。有争议的文章与丽莎的论文特点一致，而与保罗的不一致。保罗和丽莎的早期论文表现出一致性但却有不同的风格，至少在由单词计数所定义的风格上是不同的，因素图是识别论文所属作者的一个有用工具。

豪–马奇博士给丽莎A+的成绩并指责保罗抄袭，他现在正忙于和保罗的律师展开漫长的官司，这无疑会使得我们优秀的统计学家朋友身无分文。不过，重要的事情是，有一个统计方法得以展现。科学再一次获胜。

6.5.3　参阅

“Who wrote the 15th book of Oz?,” by J.N.G. Binongo in *Chance*, 16, 2, 9-17.



6.6　在帕斯卡三角上播放音乐

想很快知道几率是多少？帕斯卡三角是一个简单的数字布局，能够快速且容易地计算概率。这300年来它一直有效，所以我敢打赌，它对你也有效。

统计人员最常做的事就是计算概率，概率可以对于各种情况描述预期的结果。一个简单的例子是抛硬币。试想一下，你曾被要求对抛硬币的结果下注。有两个可能的结果，正面或反面，一次抛硬币，结果无论是正面还是反面，几率都是1/2。

如果你知道得到获胜结果的不同方法数以及可能的结果数，那么数学上计算就很容易。在抛硬币的例子中，只有一种方式能获得一个获胜的结果，并且只有两种可能的结果。我们进行多次硬币翻转，如果想要知道所有可能的结果数，以及这些组合有多少符合我们的获胜标准，那么这样的数学计算就稍微难了一点。例如，如果我想要在两次硬币翻转中连续出现两次正面，我可以列出所有可能的结果，确定使我获胜的结果数量，然后看我获胜的所有结果占据多大比例。这一比例就是获胜的几率。

但是，获胜的可能结果数往往比我们简单的掷硬币例子更复杂，因为可能有许多试验（掷骰子、购买彩票，或诸如此类）和许多不同的组合。例如，你可能想要弄清楚从一顶帽子中抽出或通过其他随机方法选择的物体中，不同元素的可能组合数量。

想象一下，你和亲戚共6人准备开车去机场，你们必须都坐一辆厢式货车过去。你并不偏好谁更多，所以你需要某个公平的方式来决定大家的位置。一同前往时，你会随机挑选两个名字坐在前排座位。



给我叔叔弗兰克的私人字条：是的，这个例子基于去年“不愉快的”感恩节。我们彼此谅解吧，至少我家这边如此，但我们认为如果你明年能把自己的车开过来就最好了。

现在，你需要知道你坐前排座位的几率，以及你可能与谁坐一起的几率。问题是计算坐前排座的亲戚有多少种不同的组合。对于这两种简单的赌注，如硬币翻转和生死攸关情况下的长途车旅行，你可以使用被称为帕斯卡三角的数字布局进行计算。

6.6.1　帕斯卡三角介绍

帕斯卡三角如图6-6所示。这种数字的布局有一些有趣的属性。下图显示它由10行组成，最底行有10个数字，但它可以扩展成无限多行变得无限大。往下走向的外边缘数都是1。临近的对角线从1开始，但是它们每往下一行，数字就增加1。

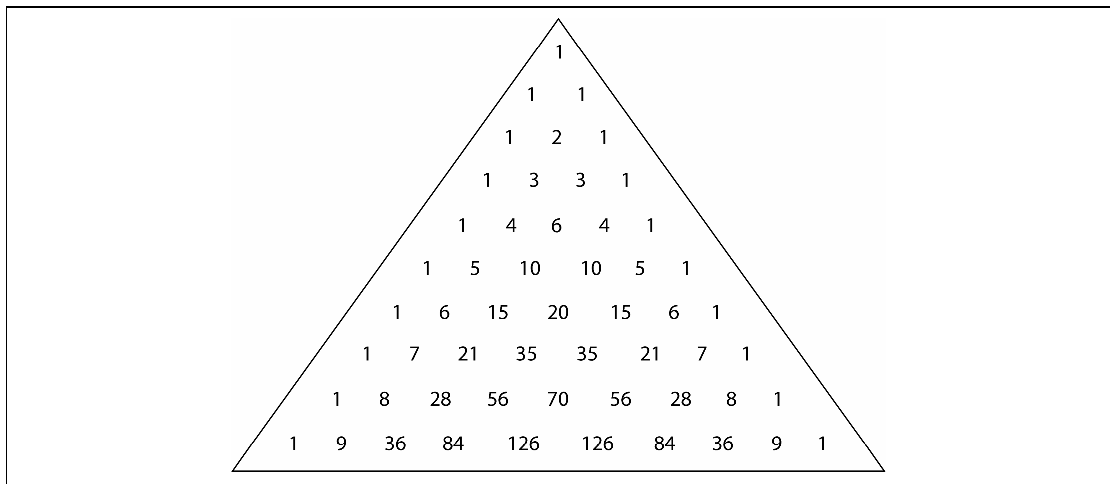


图6-6：帕斯卡三角

类似有趣的级数在整个三角中都能发现。注意，每个数字都是这个数字上面两个数字的总和：84=56+28，7=6+1，以此类推。但是，这些很酷的模式不是我们对三角感兴趣的原因。相反，我们要用它来计算各种结果的概率。

6.6.2　使用帕斯卡三角计算概率

因布莱兹·帕斯卡（Blaise Pascal，一个生活在17世纪，很聪明的早期概率理论贡献者）而命名的帕斯卡三角，已经利用了我们回答各种各样的问题所需要的计算。



虽然这种数字模式被称为**帕斯卡三角**，但发明者却不是帕斯卡，他本人也从未声称发明了它。帕斯卡的老师提出过类似的数字模式，同时期也有其他人在论著里提及。

存在一个通用公式可以确定特定类型的可能结果数。这个公式适用于恰好有两种可能结果时，因此，术语二项式系数就用来描述该公式的结果（二项式是指具有两个名字，或在统计意义上，具有两个结果）。为了确定给定数量的试验中可能结果的二项组合，可用这个计算公式：

$$\text{可能的获胜组合数} = \frac{k!}{k!(n-k)!}$$

能带入这个公式的可能范围值是帕斯卡图形上的坐标。方程中的n表示试验或事件的次数，表示在图中找哪一行。公式中的k告诉我们这一行的具体条目。沿着图形左边的那些数字1像是一个边界：它们算作0。因此，要使用三角，我们以0开始计数。



在这个公式中，一些数字后的感叹号表示阶乘，这反过来又意味着，你应该从这个数字倒计到1，并乘以所有倒计的这些数。例如，5的阶乘是5×4×3×2×1，即120。顺便说一下，根据规则，0!为1。

1. 评估翻转硬币结果的概率

下一步，我们解决稍微复杂一点的抛硬币问题，抛一枚硬币两次正好出现两次正面的几率可以使用三角来计算。

(1) 要找的那行由我们翻转硬币的数量决定：2。我们要数的那行的条目由我们想看到的正面的结果决定：2。对于我们掷硬币的例子，2次试验中出现2次正面，往下数两行至如下行。

1 2 1

(2) 然后，数两个条目到1。我们的答案是1，所以我们得到两个正面的机会有1次。

(3) 但是多少次机会里有1次呢？把你那行的数字相加，就得到了那个答案。1+2+1=4，所以我们的几率是1/4或25%。

三角也可以回答更复杂的问题。假设你想在6次硬币翻转中正好得到3个正面。

(1) 往下数6行（记得把三角顶部记为0）。你到了如下这行。

1 6 15 20 15 6 1

(2) 数3个数字你得到了20。6次硬币翻转，正好得到3个正面的不同方式有20种。

(3) 你会问，是多少可能性中的20种？对该行所有的值加和我们得到了64。64次中有20次，你会恰好得到3个正面（或3个反面）。概率大约是31%。

2. 评估一趟糟糕的自驾之旅的概率

另一种使用三角的方法是，看看以一定方式抽取的某种数量的元素有多少种可能组合。我们自驾之旅的例子关注从6人中抽取2人有多少可能组合。

一组中有6个要素，你将抽出其中2个并将其匹配起来。对于这个的问题，以及定义三角的二项式公式，把6个亲戚想象成n，要抽取的2个名字为k。

(1) 往下数6行然后跨过2个条目，你得到数字15。从6个人中抽取2个人，有15种可能的组合。

(2) 在这种情况下，你只对和某个特定的人坐在车前排的几率感兴趣。这是15种可能的组合中前排乘客的1种组合。因此，你和你讨厌的弗兰克叔叔或蒂莉婶婶，或任何人，坐在前排座椅上的概率仅有1/15。

6.6.3　生效原理

如果你真的利用二项式公式进行数学计算，那么三角中的数字会匹配你数学计算得到的值，但你会发现，三角还能回答其他问题。数字的模式、它们的级数，与其他确定概率时使用的公式都是一致的。

举例来说，6次硬币翻转总的可能翻转组合数，由累加三角中第六行的值来回答：64。你可以通过应用反转硬币后可能结果的通用公式来对其进行数学推导，求出这个值：$2^{\text{翻转次数}}=2^6=64$。

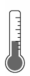
至于你既会被选中为6人中的2人，同时和你一起坐前排的人又是其他特定人的其中一个（我们去机场的例子）的几率，三角表示是1/15。但你也可以通过下面的方式计算：

(1) 成为6人中选出的两人之一的几率=2/6=0.33；

(2) 从5个“其他”人中选出1个特定的人的几率=1/5=0.20；

(3) 两者同时发生的几率=0.33 × 0.20=0.066，或1/15。

所以，当你涉及的组合和排序看起来很复杂，有如此多的可能性使你头晕时，让帕斯卡三角舒缓的音乐给你混乱的大脑带来安宁。



HACK #67

6.7　控制随想

我们的内心思想本就漫无边际，人们认为这能创造不可预测的随机路径。你可以提高周围人聚焦在你希望的事物上的概率，利用这个误解来猜测周围人的想法。

我们对令人毛骨悚然的场面并不陌生，埃德加·爱伦·坡（Edgar Allen Poe）在《莫格街谋杀案》（*Murders in the Rue Morgue*）中提到了这点：

> 我们两人陷入沉思，至少15分钟内谁都没说一个字。突然，杜宾蹦出一句话：“他是一个非常小的家伙，这是真的，能为杂技团做得更好。”“那是毫无疑问的，”我不知不觉地回答说……“杜宾，”我严肃地说，“这我无法理解。我毫不犹豫地说我很惊讶，而且几乎无法相信我的感觉。你怎么可能知道我在想（什么）？”

你是否曾经一直和某人交谈，中间你的思想开了一小会儿差？然后，你提出了你想的东西，你惊奇地发现，另一个人也在想同样的事情！

为什么会这样呢？你能做到这一点吗？你能预测另外一个人要说什么吗？是的，很有可能，

你有时可以做到这一点，有时你可以预测另外一个人要说的话。如果你们两个人都有共同的背景经历，尤是如此。

6.7.1 思想控制

我们的记忆中充满了单词、想法、故事等，它们和其他的单词、想法和故事相关联。如果你希望某人思考某一话题，这样你就可以读懂他的心思，那么，让他思考到你希望让他思考的东西，最容易的方式是提出一个与所需话题密切相关的话题。

例如，如果你希望你的朋友开始思考狮子、老虎和熊，你可以让与这一主题相关的单词进入他的思维过程，如绿野仙踪、桃乐茜、托托，甚至条纹，因为条纹和老虎是高度关联的。

所有的词在书面语和口头语中都有一定的出现频率。有些单词有非常高的出现频率（比如：这个、它，等等），而有些单词具有非常低的出现频率（比如：土豚）。此外，有些单词和其他单词组合出现的频繁相当高（比如：盐和胡椒，韵律和蓝调）。事实上，有些词经常一同出现，以至于研究发现：即使只说出一个单词，人们也会立刻想到其他的词。

通过学习这些关联，我们可以更快速地处理传入的信息。如果我们听到盐，就能想到盐和胡椒，那我们就领先一步了，并可以在晚餐同伴要求我们递给他之前就开始伸手去拿盐和胡椒。

所以，如果你想“控制”人的思想，关键是简单地知道哪些事情最经常一起发生。一个单词出现的越频繁，某人想到它的可能性越大。同样，两个词一起出现的越频繁，当只说出一个词时就想到两个词的可能性越大。

6.7.2 概率与单词联想

多年来，感兴趣的研究者已经收集了很多关于关联单词的数据，以分析对我们人类来说什么是常见的。精神科医生使用单词之间自由联想的典型知识作为读取潜意识的工具。认知心理学家使用相同的信息来映射大脑处理信息的方式。

现在已有大量的、关于线索（提出的这个词可能会导致联想）和靶子（在线索提出后想到的单词）的已知信息。表6-16展示了单词线索的样本，以及正常人会想到特别靶子的概率，比如你的朋友。该表提供了一系列的好线索和坏线索，以供你理解大部分思维是如何工作的。

表6-16：单词联想几率

线索	靶子	概率
安全套	性	0.53
颠簸的	性	0.01
西兰花	绿色	0.25

（续）

线索	靶子	概率
西兰花	毛	0.01
睡衣裤	睡觉	0.36
事故	车辆	0.36
事故	哎呀	0.01
妈妈	爸爸	0.60
妈妈	鹅	0.02
牙医	牙齿	0.42
英雄	超人	0.17
英雄	蝙蝠侠	0.02
统计	数字	0.26
统计	无聊	0.03
凉拌卷心菜	鱼	0.01

当你想让你的被试去思考某些词或想法时，像上面这样的信息就是有用的。例如，对于“性”这个词来说，“安全套”这个线索比“颠簸的”会让你有更多的幸运。



表6-16抽取自http://w3.usf.edu/FreeAssociation/网站，上面有详尽的数千单词的联想清单，由南佛罗里达大学和堪萨斯州的研究人员尼尔森、麦克沃伊和施雷伯提供。

6.7.3　建立单词联想列表

联想的观念和单词形成了人与人之间稍有差异的连接网络，但有共同文化（流行音乐或其他）或有共同经验的人，他们的连接网络是相似的。想要大声说出朋友的想法（吓死他们），你需要知道你的隐喻世界里各个角落的可能联想。

你可以进行一个小的研究，以确定对你朋友而言哪些单词彼此间有最强的关联性。创建几个有代表性的朋友或家人的样本。制造一张测试单词列表询问你的被试者，当你说出每个单词时，他们最先想到的第一件事是什么。常用短语或标题中的单词效果最好。但在随后的实际对话中，笑话、电影或歌曲里能引发兴趣的单词是最适合使用的单词类型。



现实世界里，认知心理学家在他们的研究中用数据来更多地了解思维过程，你的小型研究是一种快速方法，能够获得具有相同类型数据的小样本。

对于某一单词，如果你的很多朋友都给予同一单词响应的话，你可以假设这个响应单词与测试单词的关联性非常强。你希望将最高概率的单词组成心理泵吸向预测结果。

6.7.4 生效原理

人类的大脑是如此高效，以至于只要单词或概念已经深入学习过，它就能对其进行处理。研究发现，当人们被要求说出一连串的字母是否能组成一个单词时，他们会对任务之前展示给他们的、预先学习或激活的单词作出更迅速的回应。例如，如果展示了条纹的一词，然后展示老虎或柠檬单词，相比柠檬，人们会更快地回应老虎。

谈论与其他单词或主题密切相关的单词或主题时，你的大脑的思考过程和朋友的一样，神经元的激活扩散到大致同一时间被唤醒的神经元。你的大脑已经习得，某些单词和话题几乎总是同时出现，所以当联想单词或主题之一被激活时，大脑中与被激活单词和主题相关联的区域也应该被唤醒。这样一来，你的思维过程得以顺利进行。

6.7.5 其他生效领域

这种特殊的思维伎俩有一些失败的风险，尤其当你使用低概率的关联时。然而，你可能只是享受自己偷偷操纵别人的感觉，并不想通过它做大秀。

我们可以激发人们做很多看起来如同自然发生的事情，因为这些事情的发生是如此频繁且毫不费力。例如，只需通过自己打哈欠就可以让别人打哈欠，这很有可能。你甚至可以通过谈论打哈欠或谈论睡眠让朋友打哈欠。（事实上，我写到这里时，就打了个哈欠。）同样地，如果有一些食物，你想将其作为正餐，那么你可以通过提及这种食物让你的家人也渴望它。

你可能已经自我激发很多次了。当你正听着喜欢的CD时，一首歌结束了，你是不是在下一首歌开始播放前已经开始在脑海里听到这首歌了？如果你知道某人会把什么事物关联在一起，那么在你激发这些事物之后，预测此人的想法会变得相对容易。这是结婚的人往往能接续彼此话语的一部分原因。

6.7.6 无效领域

如果某人和你的语言背景不同，他们讲不同的语言或讲不同的方言，那么他们可能和你有没有相同的联想词。

但如果一个单词有若干个可能性相同的联想词，这也无效。例如，如果你用热（hot）这个词激发别人，最开始有些人可能会想到天气（热hot，冷cold）。有些人可能会想到食物（热狗，hot dog），也有人可能会想到他们仰慕的人（性感姑娘，hot babe）。

当你看到热这个词时，你觉得你会想到哪个词？我就知道你会这么说！

——吉尔·罗米尔和布鲁斯·弗雷



6.8　搜索超感官知觉（ESP）

虽然大多数科学家都认为没有太多证据表明ESP确实存在，但科学家可能是错误的。你、你的朋友或你的猴子可能就有ESP，事不宜迟，现在我们就去找出来！

超感官知觉（Extra-Sensory Perception，ESP）一词用来形容独立于传统5种感官的感知，传统的5种感知是：视觉、听觉、触觉、味觉和嗅觉。最先使用这个词的是20世纪二三十年代就职于杜克大学的心理学家J. B. 莱茵（J.B. Rhine）。那个时候有很多令人激动的事，因为莱茵和他的同事们能够识别出似乎具有ESP能力的个人。在那个时期至70年代的大众传媒和一些科学论文中，有人甚至理所当然地认为，存在ESP这样的事，我们都在一定程度上拥有该特质。

但是，时至今日，你真的没有听到太多关于ESP的消息，大多数科学家已经得出结论：这样的事情可能并不存在。更具体地说，它还没有达到科学验收的标准，即一些假设没有满足期望的标准，如实验证明，复制研究，等等。但是，你可以添加数据并开展自己的研究，确定你或你的朋友是否具有超自然的能力。

6.8.1　识别超自然能力

虽然有各种所谓的超自然能力，从读心术到用意念移动物体，但研究ESP的传统方式一直是使用一副叫齐纳卡的扑克牌。齐纳卡有25张相同背面的卡片。每张卡面显示5个符号中的1个：圆形、十字形、方形、星形或波浪线，如图6-7所示。

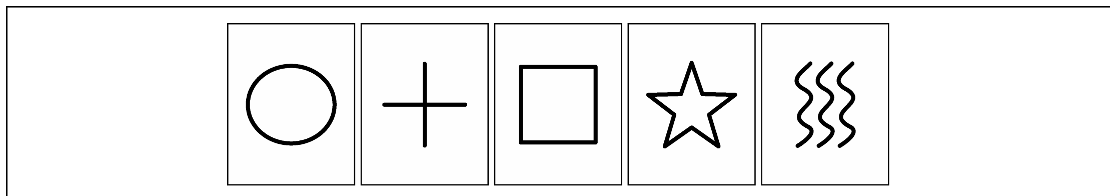


图6-7：齐纳卡

即使你手头没有这些卡片，你也可以很容易地用一包空白卡片和一支黑色马克笔制作出它们。只要确保没有人可以看穿它们即可（除非他们有超自然能力，在这种情况下，他们也能看穿你）。每个符号制作5张牌，共25张。

使用洗好的齐纳卡进行ESP测试，有以下几种不同的方式。

- 在卡牌被翻转前，一个人尝试按顺序报出牌面。
- 一个人看每张牌的牌面，尝试将其以心灵感应的方式“发送”给坐在旁边的人。
- 一个人在另一个房间或在一个遥远的位置，看每张牌的牌面，尝试将其以心灵感应方式发送给远距离的另一个人。有时候，接收者想象他们和发送者在同一房间里，可以看到牌。

不管你选择什么方法，流程是要遍历25张牌并跟踪命中数和未命中数。被试正确识别了25张卡牌中的多少张？在一些研究中，接收者有时在遍历全部卡牌的过程中就会被告知他们的表现如何，有时直到实验结束才会被告知表现如何。结果变量是被正确识别的卡牌数量或百分比。



在ESP的研究中，试图读取别人想法的人是**接收者**，想让自己的想法被读取的人是**发送者**。

6.8.2　分析结果

如果结果是仅由几率导致的预期结果，那么就将这一结果作为被试没有超自然能力的证据。如果被试答对的数目比仅靠猜测答对的数目多得多，那么这一结果有望表明被试可能有ESP。

那么，什么是几率导致的预期？如果你猜25张牌，同一类型的牌各有5张，那么仅靠几率约能猜中5%。例如，想象一下，25次中的每一次你都猜星形。你保证会得到5次命中和20次未命中，因为你知道总体来说星形出现的次数正好是5次。如果你每次随机猜5种可能中的1种，那你的平均命中率也将是5/25或20%。

但是如果你有比20%更高的成功率，说明什么？如果你25次中正确识别了6次，成功率是24%呢？我们是否应该把它视作一种证据，证明有几率之外的事物在发挥作用？我们需要对不同可能结果进行统计分析，以确定达到多大比例时应该视为如此不同寻常，以至于它一定是某种不同寻常的东西存在的证据。



统计检验只揭示几率是否是对结果的最好解释。对于我们的实验，统计显著的成果并不能证明ESP存在，只能证明几率不是最好的解释。毕竟，对于高命中率的最好解释可能是接收者从发送者的眼镜里看到了被反射的牌，或其他不太有趣的原因。

我们知道，在短期内（或用统计术语表述，在一个小样本内），结果和总体不同是常见的。但是，我们也知道，和总体值之间大的差异是罕见的，尤其是从长远来看（或大样本）。事实上，发现给定大小的样本值和总体值之间存在差异的概率和样本的大小有直接关系。

对于ESP实验，样本量是猜测或试验的次数，总体是所有试验中不同符号的已知分布。对于总体值，任何次数的猜测正确率都是20%，这就是几率导致的预期。如果在样本值和总体值之间存在巨大差异，那么可能就有几率之外的事物在起作用。

适用于这里的统计分析称作Z检验，Z检验用来比较观测比例和预期比例。它和其他常见的统计检验类似，如t检验[Hack #17]，t检验计算差值，并判断如果一个给定的样本确实是从有某些特征的总体中随机抽取的，这样的差异被发现的频率。

任何差异的概率都取决于样本的大小。例如，如果25次尝试，一个人猜对了24%，而不是预期的20%，那么对于这个分析，需要的信息是：

- 样本量25；
- 0.24的观测比例；
- 0.20的预期比例。

不展示这个特殊分析的公式和计算过程，我会把结果告诉你。对于25次猜测，只凭几率，被试有31%的可能至少正确猜中24%的牌。另一种表述方法是：100个参加研究的被试，其中有31个人会得到这样的结果或比这更好的结果。因此，24%的命中率高于平均水平，但还不至于不同寻常到让我现在就将这一情况上报国家。

如果你试验超过25次，那么命中率如何？表6-17展示了正确猜中给定百分比（或更高百分比）的几率。此表假设预期命中率为20%。

表6-17：选定的ESP命中率的可能性

猜测数	正确百分比（命中率）	达到或超过命中率的可能性
25	20%	50%
25	30%	11%
25	40%	1%
25	50%	0.01%
100	20%	50%
100	30%	1%
100	40%	0.000 01%
100	50%	0.000 000 000 001%

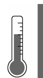
注意，随着样本容量增大，极端结果的可能性大幅下降。例如，只有25次猜测时，获得40%命中率的几率约为1%；如果你遍历25张牌100次，你很可能只有1次会做得那么好或更好。但是，如果你猜100次，也许遍历25张卡牌4次，你得到40%或更高正确率的几率仅有1/100 000 000 000 000！

6.8.3　多少才够

如果你想进行ESP实验，你应该建立一个标准：一个现象的不可能性必须达到什么水平，你才考虑将其作为有表面几率之外的东西在发挥作用的证据。通常情况下，在统计研究中，如果结果偶然发生的几率为5%或更小，其结果就被视为统计显著。对于有25张齐纳卡和25次猜测的ESP实验，你猜对8个或更多卡的概率约为7%。你猜对9个或更多卡的几率只有2%。因此，某个介于8至9的命中标准是科学合理的。

我内心的怀疑感迫使我必须给你一个警告。如果你进行这个实验，在你自己身上或别人身上

获得了显著结果，这是很酷的。但是，如果你能重复这一发现，在同一个人身上复制实验并获得相似结果，这将使一切变得精彩！如果出现这种情况，立刻给我发电报，我会卖掉我的房子投身其中，我们将踏上名利之路！



HACK #69 6.9 治愈合选症

两个独立事件同时发生的概率永远不会比其中任何一个单独事件发生的概率更高。出人意料的是，人们常常意识不到这个常识性真理。

试想一下，在一次晚宴上，朋友将你介绍给约翰，他是一个令人愉快、身材高大且稳重的男子。你与约翰闲聊了几分钟，发现他很友好、很爱笑，但不是很聪明。约翰急于谈论目前正在进行的世界职业棒球大赛，也问你开什么车。

晚宴结束，在回家的路上，你的爱人询问晚饭前和你交谈的那个人的情况。你分享了约翰的一点信息，但意识到自己从未了解他是做什么的。事实上，正如你意识到的一样，你知道的关于他的信息真是太少了。你的爱人决定和你玩一个小智力游戏，并解释道：

> 我对约翰有所了解。我会提供一系列关于他的陈述。它们可能是真的，也可能是假的。所有陈述都可能是真的，也都可能是假的。这些陈述也可能是真假混合的。我希望你基于自己对每个陈述为真的信心大小，对它们进行排序。当我们完成后，我将诊断你是否患有一种称作合选症（Conjunctionitus）的常见脑疾病。

然后，你的爱人要求你对如下关于约翰的陈述排序，猜测哪个最有可能是真的：

(1) 约翰是计算机科学家；
(2) 约翰是汽车推销员；
(3) 约翰是前棒球运动员；
(4) 约翰是共和党人；
(5) 约翰曾经是打棒球的计算机科学家；
(6) 约翰是跑马拉松的传教士；
(7) 约翰演奏单簧管；
(8) 约翰结婚了。

像很多人一样，你可能会把陈述3（前棒球运动员）列为最有可能为真的陈述之一，把陈述1（计算机科学家）列为最不可能的。到目前为止，陈述内容还没有那么疯狂，至少它们都是基于你刚才谈话的合理猜测。

你的合选症症状，与你分配给陈述5的排名位置有关。我打赌你把它排在陈述1的前面，认为其可能性大。如果是这样，那你可能患有合选症：一种导致人们作出糟糕的概率判断的症状。

事实是，两个事件一起发生的概率永远不会大于任何单独一个事件发生的概率。因此，“约翰曾经是打棒球的计算机科学家”不会比“约翰是计算机科学家”更有可能。但是，不要害怕，在这种情况下，若要提高你作出可能性判断的能力，第一步是承认你有问题。接下来的步骤是了解状况，然后就可以开始治疗了。

6.9.1　问题

虽然更多的信息可能使一个描述看起来与某人或某事更加相似，或对某人或某事更具代表性，但是更多的信息不会使事情更有可能。如前面提到的，两个事件一起发生的概率不会比它们中的一个单独发生的可能性高。考虑一个人在这个世界上的所有可能事件。你如何决定约翰的哪些事情是最有可能的？你可以从观察*基础概率*开始。

在这个世界上，相比计算机科学家、汽车销售员、前棒球选手、共和党人、传教士、马拉松运动员和单簧管玩家，已婚男人的数量可能更多。因此，很有可能约翰已经结婚了。你把这个可能性排在哪里？

因为我们可能真的不知道所有其他可能性的*基础概率*，所以我们可以使用关于约翰的已知信息去预测哪些陈述是最有可能的。我们明确知道，如果考虑包含所有前棒球选手的群组和包含所有计算机科学家的群组，大概只有少数人同属于这两个群体。因此，在曾经打棒球的计算机科学家群体中的可能性，一定比在计算机科学家群组或在前棒球选手组的可能性要小。

但是，大多数人，即使他们是理性的、聪明的决策者，也会被拉向合选的句子（即列出两个独立“事实”的句子），仿佛将“事实”列在一起使它们更可能是真的。即使（也许尤其是）第二个“事实”本身似乎就不太可能的情况下。

6.9.2　合选连结的原理

为什么我们的头脑往往以这种方式工作？20世纪70年代，诺贝尔奖得主丹尼尔·卡尼曼和他的同事阿莫斯·特沃斯基给大学生展示了几个问题，其中一个选项高度代表一个给定的个性描述，一个选项与描述不一致，一个选项包括高度相似和不一致这两个选项。

也许最众所周知的、反映合选谬误的著名问题（至少在认知心理学界）是琳达问题：

琳达31岁，单身、直率，也很聪明。她主修哲学。作为一名学生，她深切关注歧视和社会公正问题，而且她还参加了反核示威游行。

被试需要判断下列陈述为真的可能性，并按高低顺序进行排列：

(1) 琳达是一名小学教师；
(2) 琳达在书店工作，还参加瑜伽课程；

(3) 琳达在女权运动中很活跃；

(4) 琳达是一位精神科社会工作者；

(5) 琳达是妇女选民联盟的成员；

(6) 琳达是银行柜员；

(7) 琳达是保险推销员；

(8) 琳达是银行柜员，并积极参与女权运动。

卡尼曼和特沃斯基（和许多其他曾经复制过此研究的人）发现，人们一致都把选项8（积极参与女权运动的银行柜员）视为更有可能，将其排在选项6（银行柜员）之前。这是因为选项8提供了更多的信息，看起来更能代表琳达。因为我们期望她在政治上活跃，但我们不指望她是一个银行柜员，她看起来会是银行柜员的唯一途径是：她也积极参与政治活动。

然而，我们知道，选项8不会比选项3或选项6的可能性大，因为如果我们想象所有活跃在女权运动中的人，他们（也许是一个小的子集）的一个子集是银行柜员。同样，如果我们想象世界上所有的银行柜员，一个子集（同样，也许是一个小的子集）活跃在女权运动中。因此，作为一个银行柜员的可能性要大于作为活跃在女权运动的银行柜员的可能性。有道理吧，对不对？但你的思维不以这种方式运转。



两个事件一起发生的概率无法大于其一发生的概率，这被称为**合选规则**。很多人往往认为合选的两个事件有时会比单独一个事件发生的可能性大，这种事实被称为**合选谬误**。

6.9.3　治愈

停止错误地思考这类命题，治愈方法很简单：

(1) 别说了；

(2) 停下来；

(3) 不要那样做。

合选谬误可以在工作的许多地方看到。注意它可能发生的情境，并分析该情境。例如，你可以向一个棒球迷询问他喜爱的且经常打不出本垒打的球员的情况。询问他该球员是否在接下来的比赛更有可能作出如下哪件事情：

- ❑ 打出一个本垒打；
- ❑ 出局；
- ❑ 出局并打出一个本垒打。

粉丝可能认为，在比赛中，一个本垒打加一个出局比仅仅一个本垒打的可能性要大。但事实

不是这样的。



有一些情况下，选择合选命题可能也可以。如果两件事情必须一起出现（如电闪雷鸣），那么两者都发生的可能性和它们其中之一发生的可能性一样。如果你增加关于雷鸣和闪电的陈述，并对比雷鸣（没有闪电）与电闪雷鸣的可能性，那么，其实，电闪雷鸣的可能性会更高。然而，这只适用于如果没有另一个，这一个也永远不会发生的情况。

一旦意识到这个概率估计的常见错误，你会发现它无处不在。例如，你可以很容易地在政治预测舞台上找到合选谬误。乔治・W. 布什更倾向于：

- 提名一位温和的最高法院法官；
- 提名一位温和的最高法院法官和一个右翼最高法院法官。

当然，你现在知道答案了，但许多政治分析家可能会和你争论，那是因为他们有病。他们有合选症。你曾经也有病，但现在治愈了。

6.9.4 参阅

- Tversky, A. (1977). “Features of similarity.” *Psychological Review*, 84, 327-352.
- Tversky, A. and Kahneman, D. (1974). “udgment under uncertainty: Heuristics and biases.” *ience*, 185, 1124-1131.

——吉尔・罗米尔



6.10 用 Etaoin Shrdlu[3]破解密码

你从来都不知道什么时候你将不得不破译一个密码，不论它来自男神詹姆斯・邦德的拦截消息，还是你的医生字迹潦草难以辨认的处方笺。以下是你所需的所有统计技巧，代号为003.14159。

你可能已经注意到，你电脑键盘上的某些键比其他键在更短的时间内变脏或磨损。那是因为你敲击它们的次数比其他键盘更多。你可能还注意到，这些字母往往在键盘中间，或者更确切地说，当你的手放在键盘正中时，它们在你手附近的小圆圈里。

磨损的按键及其在标准打字机中的位置（又名QWERTY，顶行前6个字母）都是基于它们在英语中的使用频率而定的。字母表中的不同字母在语言拼写中的使用频率是不同的。通过运用这

注3：Etaoin Shrdlu是本行作废的意思，此处指用这几个字母来破解密码。——译者注

些字母的已知频率及其他统计技巧，你可以快速解码机密文件，无论它们是达芬奇的日记、报纸上的谜题，还是在电视上被Vanna White翻转的、大而明亮的字母。

6.10.1 单替换密码

最简单且最古老的以字母为基础的代码类型是单替换形式。在这些代码中，一些消息单词的实际字母被转化为字母表中的其他字母。用这种方式编码的最简单形式是，整个消息中的相同字母被替换为同一个字母。例如，一个简单的密码文可以使用表6-18所示的替代方式，其中上面一行的字母（原文本）被底部一行（密码文本）的字母取代。

表6-18：单替换密码

原文本	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
密码文本	N	A	O	B	P	C	Q	D	R	E	S	F	T	G	U	H	V	I	W	J	X	K	Y	L	Z	M

有了表6-18这样的代码，以下的原文本段落：

Tom appeared on the sidewalk with a bucket of whitewash and long-handled brush.

以密码文本表示就是这样的：

Jut nhhpnipb ug jdp wrbpynfs yrjd n axospj uc ydrjp yhwd ngb u fugq-dngbfpb aixwd.

这段话看起来毫无意义，但有了表6-18所示的线索，任何人都可以轻松地把无意义的字母替换为原来的字母，这样《汤姆·索亚历险记》（*Tom Sawyer*）第2章第2段的开头就显示出来。

6.10.2 用概率来解码替换密码

当然，破译密码时，真正的任务是没有代码线索的。现实生活中的代码破解人员和幸运之轮的获奖选手使用相同的工具来解决他们的问题：他们运用英语单词中字母的已知分布。

电脑、电脑分析和数百万电子书籍的出现，已经使计算字母表中的每个字母的确切概率成为可能，虽然密码学家（代码制造者和破解者）已经知道这些基础知识很长时间了。下面是其中的一些基本知识：

- 英语中，最常使用的字母是E；
- 最不常使用的字母是Z；
- 最常用的辅音是T；
- J和X与很少被用到，Q也一样；
- 当Q被使用时，几乎总是伴随着U；

- 在英语中，只有A和I作为单个字母构成的单词来使用。

哪怕只掌握这些基础概率事实，你也可以开始着手对一个密码进行解码，如我们的马克·吐温段落。在乱码版本中，最常出现的字母是P和N。因为N是一个单一字母的单词，所以它不能是E（N最有可能是A），所以对P替代字母的最优先猜测是E。

仅靠字母分布的一点点知识，我们已经确定了E和A的替代字母。我们无法肯定这是正确的，但像任何优秀的统计学家一样，我们认为自己可能是正确的。表6-19显示了字母表中每个字母的可能分布。

表6-19：英语中字母的频率分布

字　母	频　率
A	8.04%
B	1.54%
C	3.06%
D	3.99%
E	12.51%
F	2.30%
G	1.96%
H	5.49%
I	7.26%
J	0.16%
K	0.67%
L	4.14%
M	2.53%
N	7.09%
O	7.60%
P	2.00%
Q	0.11%
R	6.12%
S	6.54%
T	9.25%
U	2.71%
V	0.99%
W	1.92%
X	0.19%
Y	1.73%
Z	0.09%

6.10.3 ETAOIN SHRDLU

奇怪的短语“ETAOIN SHRDLU”是一种帮助我们记住最频繁出现字母的记忆口诀（记忆方法）。这12个字母占全部字母出现频率的80%以上。

你可能已经注意到，在ETAOIN SHRDLU中，字母顺序并不完全与表6-19所示排名一致。但顺序足够接近，并且读起来比完全正确的排序更容易。另一件要记住的事是，任何“最终的”字母概率列表取决于字母计数的来源材料。你可以找到许多不同的字母排序和频率列表，其中一些和其他的稍有不同。

例如，一个制作英语文本中字母使用统计分布列表的组织，其结论来源于对7本文学名著的计算机分析以及实际的字母出现次数，如《简爱》（*Jane Eyre*）和《呼啸山庄》（*Wuthering Heights*）。7本书中两本是关于人猿泰山（Tarzan）的小说。我猜，如果我们比较这张表和其他表的字母分布，我们会发现，这张表显示的字母Z出现的比例大于使用其他来源的表。但是，对于常见的字母，比如E、T和A，对于它们作为密码破译的首选猜测字母，人们已经形成广泛的共识。

幸运之轮策略

电视节目真人秀《幸运之轮》（*Wheel of Fortune*）中，在最后解决大难题之前，友好的制片人会提供某些字母，并显示字母是否出现在刽子手式的短语中。他们提供R、S、T、L、N和E。当然，给定这些字母是因为它们非常常见，并在我们的前12名中：ETAOIN SHRDLU。玩家被允许再选择3个辅音和另外1个元音。利用我们的字母频率统计知识，一个较好的基本策略是：选择A作为元音，并选择H、D和C这3个最常见且尚未出现的辅音。

6.10.4 编码文本的统计分析

下面是如何使用这些字母统计量在现实生活中解码秘密消息或解决一个难题。如果编码的文本很长，这方法效果最好，但是它对较短一些的段落发挥出效果也足以令人吃惊。计算编码的、替换的字母（密码文本）分布，然后把它和表6-19所示的分布进行对比。

图6-8使这个方法的表述更为形象。该图只展示了前10个最常见的字母，但分析中会使用所有的字母。这个例子假设表6-18所示的编码文本和替代密码被大量使用。

因为最常见的替代字母为P，其次为J，所以破解代码时，首先猜测P是否真的代表E，以及J是否真的代表T。可以沿直线往下逐个猜测每一字母。从最频繁出现的字母开始，向列表下方移动，一名密码破解人员可以很快看出这些猜测是否正确，他们不断改变猜测，直到英语单词开始出现。

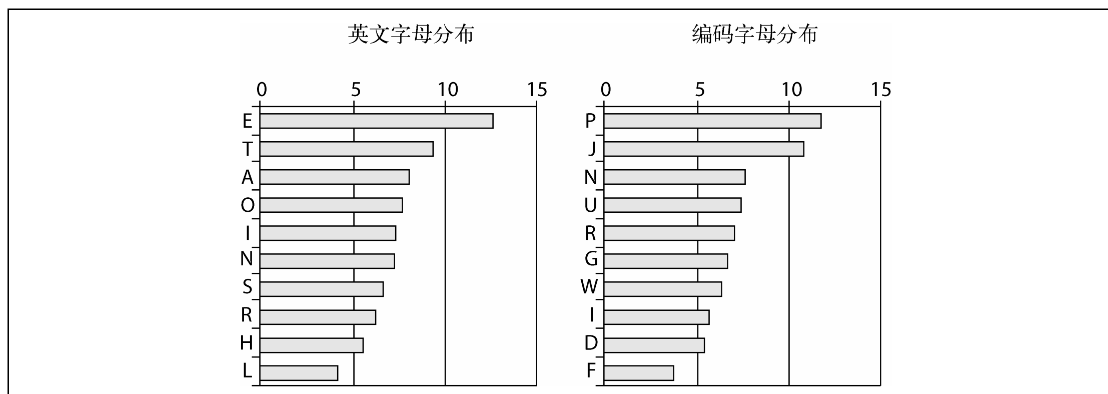


图6-8：英文字母频率（左）和编码字母频率（右）

6.10.5 其他常见的字母模式

除了知道个别字母的出现频率，优秀的密码破译者还使用其他的字母模式信息。

- 单词最有可能以T、O、A、W或B开头。
- 大多数的单词以E、T、D或S结尾。
- 如果某个字母在单词里连续出现两次，它们最有可能是SS、EE、TT、FF或LL。
- 频繁出现的两字母单词包括of、to、in和is。
- 到目前为止，最常见的三字母单词是the和and；其他较常见的三字母单词包括for、are和but。
- 往往成对出现的字母包括TH、HE、AN、IN和ER。
- 最常使用的单词是the、of、and、to、in、a、is、that、be和it。
- 也许还能指出哪些单词是人们常写的，在书面文本中，最常用的前100个词包括dollars、great、general和public。debts差一点就能进入前100名，但它也很常见。

6.10.6 参阅

- 一个很好的关于单替换密码的解释可以在http://en.wikipedia.org/wiki/Frequency_analysis网站的频率分析条目中找到。
- 本Hack涉及的一些统计数据可以在http://www.data-compression.com和http://www.scottbryce.com找到。你还可以在这个网站上找到关于如何使用统计来解决密码电文和其他代码的有用信息和建议。



6.11 发现一个新物种

虽然每天都有物种走向灭绝，但偶尔还是会发现未知的新物种。出人意料的是，利用统计方法而不是生物方法，可以达到鉴别物种的目的。

几年前，一个新的负鼠物种被确认。这个新物种被命名为trichosurus cunninghamii。Trichosurus代表，嗯……负鼠（我猜的），cunninghamii部分代表它的发现者，罗丝·坎宁汉（Ross Cunningham），一位澳大利亚国立大学的统计学家。如果你想有一个因你命名的物种，可以接受统计提供的帮助。

6.11.1 用统计鉴定物种

有一大族系的统计分析方法，它们着眼于一堆变量并发现变量中自然出现的分组。通常来说，变量分组或集群的鉴定依据它们之间的相关性[Hack #11]。

有一种使用这种策略的方法，它试图找出相关的维度，或无形的、能解释一堆不太重要变量的大型基础变量。这种方法就是*因素分析*，我们已经在其他章节看到它如何识别作家的写作风格[Hack #65]，除此之外，它还有其他用途。

统计学充满了类似的技术，可以识别出维度、根本原因，还有分组。对于有生物倾向的、愿意识别新物种的统计学家来说，确定分组的目标是非常有用的。

从技术上来说，如果某组动物属于一个独立的物种，那它们必须共享唯一的一组生物特点，使其区别于同类动物。当然，同一科属的动物都长得略有不同，但另一方面，人和人之间有很多的不同，但我们都是同一物种（我的叔叔弗兰克的存在或许能证明这一规则也有例外）。

如果一组动物，如坎宁汉博士的负鼠，它们之间的共同点比与同科属的其他生物的共同点要多，那它们有权将自己看做一个候选新物种。统计可以确定“它们彼此之间更相像，与其他物种的差异比仅靠几率产生的更多”的临界点在哪里。

将坎宁汉的发现作为一种模型，你要实现自己的发现，下面有几个步骤要遵循。

1. 收集数据

这种负鼠已经在澳大利亚人的眼皮底下存在200年以上，但没有人注意到。实话实说，它看上去非常像其他的负鼠，其中最常见的是trichosurus caninus，现在叫短耳负鼠（short-eared possum）。

有一段时间，人们认为这些小家伙真的只有一个品种。坎宁汉博士的一部分工作是收集和整理他周围野生动物的描述性数据。因此，他有大量的、关于各种负鼠身体各部分的、非常具体的定量描述，眼睛、耳朵、鼻子和喉咙，还有其他的物理测量。

2. 选择统计方法

坎宁汉选择了一种和因素分析相似的技术，但它有一个更具气势的名称：*典型变量分析*（canonical variate analysis）。你可以采用使用变异分数的任何方法来创建不同的组。其中一些在本书中有讨论，比如因素分析，本Hack之前的章节也提到了很多其他有效的方法。



如果你真的是擅于统计的人，那么知道**典型变量分析**和**判别分析**（discriminant analysis）及**多元方差分析**（or multivariate analysis of variance，MANOVA）具有功能上的一致性会对你有益，判别分析和多元方差分析是另外两个用于创建线性复合变量的方法，目的是定义两个或两个以上截然不同的群体。

坎宁汉用这种统计方法检验这个假定的单一物种（你知道的，就是trichosurus caninus负鼠）的描述性数据，并证明它们可能是两个不同的物种。

3. 选择一个假设并分析数据

统计学家检验假设，所以你应该在开始分析时就作出这样的猜测：提供给你数据的族群之间是否存在区别。

在我们的成功案例中，坎宁汉假设数据来自两个不同的物种群体。那么，该方法（当然，用计算机进行计算）可以确定哪些变量作为理论组之间的主要区别特征效果最好。



使用典型变量分析和其他类似回归工具的区别是，当在回归分析中使用变量进行预测时，研究者有一些关于实际科属分数的已知数据，即它们属于[Hack #13]哪个“组”。这里的方法是在不知道正确答案的情况下摸索地进行。相反，它可以找出与手头的变量最不同的群组。

下面是坎宁汉使用的变量：

- 头长；
- 头骨宽度；
- 眼睛大小；
- 耳长；
- 体长（从鼻子到卷曲的尾巴的尖端）；
- 尾长；
- 胸宽；
- 足长。

虽然还考虑了其他变量，但坎宁汉选择这些特征是因为最终发现它们是区分不同物种的最重要的方面，并且这些特征可能不受环境影响。

4. 解释结果

任何统计分析的最后一步都是描述和理解你的发现。对于发现新物种来说，你需能足够详细地描述新物种，以把它和其他同类物种区分开来。

坎宁汉使用的方法确定了由不同权重的生物变量组成的一系列方程，目的是找到最能识别两个不同组群的组合。这些方程（该方法将其称作变量）类似于回归方程，结果或标准变量用来确定负鼠属于哪个组。

下面是一个最好的公式，能够解释数据库中高达89%的负鼠特征差异：

$$（头长 \times 0.44）+（头骨宽度 \times 0.07）+（眼睛大小 \times 0.05）+（耳长 \times 0.82）+（体长 \times 0.35）+（尾长 \times 0.72）+（胸宽 \times 0.16）+（足长 \times 0.70）$$

我已经提供了研究中标准化的权重，因此我们可以将它们进行相互比较。最大的权重代表负鼠的这个身体部位在数学选择的两组负鼠间最为不同。

在这个公式中，你会发现两组负鼠的耳长、尾长和足长最不相同。从统计学上来说，变异的解释量是如此之大，以至于坎宁汉认为数学上确定的这种分组是真实的。从数据中发现的两组负鼠实际上是两个不同的负鼠物种，而这个物种可以通过它们的耳长和其他几个变量来定义。前面所示公式的权重越大，两个物种在这些身体部位上的差异就越大。

6.11.2　两个负鼠物种

表6-20显示了由我们的统计学家和他的数学首次确认的两个负鼠物种的官方描述。注意，它们甚至都是基于统计分析中发现的关键预测变量来命名的！

表6-20：两种常见的澳洲负鼠

	trichosurus caninus	trichosurus cunninghamii
通用名称	短耳负鼠	山刷尾负鼠
居住地	北方	南方
耳朵	短耳	长耳
足	小足	大足
头	大头	小头
尾巴	长尾	短尾

那么，现在开始收集你在纱门上发现的那些奇怪的、散发恶臭的昆虫数据吧，这样你就踏上通向伟大和不朽的道路了。恶臭的昆虫是一个物种还是两个？你来告诉我。

6.11.3　参阅

我在这篇美妙的文章里第一次了解了这种鉴定物种的方法：Hall, P. (2003).*Chance*, 16, 1。



6.12　互联

“六度分隔”的概念不仅是对社区的一种新时代比喻，或代表演员凯文·贝肯举办的聚会游戏。如果你想实际地测试我们都认识某个人，而这个人认识其他所有人这一观念，找出你和大家紧密相连的程度。

我认识一个人，而他认识一个曾经为美国总统工作的家伙。世界真小，不是吗？我不是说我有令人得意的关系网，但我离这个自由世界的领袖只有两个握手的距离。在你被震惊之前，你应该知道，你离世界上几乎任何人可能只是几个联系之隔。

任何两个人都在六度分隔理论内，这可能是真实的，这个神奇且经常被引用的数字6实际上来自于一个真正的科学研究！这里有一些巧妙的研究方法，向你揭示连接我们所有人的无形关系，或者至少让你和鸡尾酒会中的另一些人有连接关系。

6.12.1　六度分隔理论

有一个作品叫《六度分隔》（*Six Degrees of Separation*），作者是约翰·格尔，威尔·史密斯主演了改编自此作品的同名电影。还有一种流行的派对益智问答游戏，有时也被称为“凯文·贝肯的六度”，即尝试通过一系列的电影和其他表演去连接任何男演员或女演员，直到他们与演员凯文·贝肯有共同的连接。

这个短语和概念来自一项对小世界问题的研究。你是否曾经在一次聚会上或一间咖啡厅里和陌生人聊天，然后发现你们都认识同样一个人？社会心理学家斯坦利·米尔格兰姆在20世纪60年代末（当时比现在有更多的鸡尾酒会）就对这种现象很好奇。社交网络中有多少关系重叠？如果我们都聚在一起，列出我们认识的每一个人，总会有某种连接吧？也许，从自身的熟人关系网的中心出发，随着我们越来越往外探索，最终我们会发现自己几乎和每个人都一定的联系。但是那需要多少连接呢？

只有一度分隔意味着我们都彼此认识。嗯，我不认识你（无意冒犯），所以我们知道，如果要连接所有人，一度分隔太少了。会不会只有两度分隔？如果我们彼此不认识对方，没准我们有一个共同的朋友？

所以问题是：在你和其他任何人之间有多少度的分隔？为了得到答案，使用本Hack的方法进行一个大的或小的研究。

6.12.2　做一个大研究

怎样才能研究我们是否真的生活在一个小世界里这个问题呢？最好的办法是复制斯坦利·米尔格兰姆使用的方法。

1. 选择一个目标

米尔格兰姆住在马萨诸塞州的波士顿，他最开始选择了一位他认识的、在本地工作的人。米尔格兰姆希望建立的最终链接的末端，不是凯文·贝肯，而是一位同意作为目标的股票经纪人。你可以挑选你最好的朋友或你所在学校的校长，或你所在大学的校长。但是，首先你要获得他们的许可（一些关于伦理的东西）。

2. 招募被试

然后，米尔格兰姆随机从两个社区抽样：波士顿和内布拉斯加州的奥马哈。采取这种抽样方案是为了代表任意一个人认识目标的可能性的两个极端。从附近的人和远距离的人开始，他们的数据平均值应该非常具有总体代表性。米尔格兰姆使用了300个随机选择的被试人员。你应该在时间和花费的允许范围内招募尽可能多的人。

3. 训练被试

米尔格兰姆以邮件形式给每个被试人员邮寄了一个小包裹。该包裹包含了研究说明以及一封给波士顿经纪人的信。说明要求他们把那封信交给我们的股票经纪人，但只有当他们直接认识他时才能把那封信给他。如果他们不直接认识他，就被要求记录一些信息，如他们的名字，并把这个包裹寄给某个他们自己认识的、更有可能认识经纪人的人。那些在链接里的下一波人收到了同样的带有说明和信的包裹。如果他们认识经纪人，那么他们可能已经把信交给了经纪人，或者把它寄给了链接中的第三链接，等等。

在你自己的研究中，一定要明白、清楚地编写说明，还有，现在你可能要为这个研究的合法性作出解释，告诉大家这不是一个商业游说，也不是连锁信（我猜尽管它的字面意思就是连锁信），所有考虑到的免责条款都对你有所帮助。如果有人质疑这个项目的合法性，你还应该附上自己的联系信息。

4. 收集和分析结果

经过一段合理的时间后，你和目标人联系，并收集所有收到的信件。在每封信里，数出形成链接的名字个数。计算所有不同长度链接的平均值，以确定典型的链接数量。找出涵盖最大链接数的最小数字，这样你就有最大的距离。

在米尔格兰姆的研究中，波士顿目标人最终收到约100封信。其中，链接的平均数是6，因此，“六度分隔”中的数字六起源于此。

但是请注意，并非所有信件都成功到达，所以我们无法从这个研究中得出6是真正的正确数字。这项研究也只是在美国进行而已，并没有在全世界范围内开展，所以地球上任意两人之间只有几度分隔的宏伟观点是基于哲学的，而不是基于经验的。



考虑到对被试的复杂要求，米尔格兰姆拥有非常高的响应率。这并不奇怪，因为米尔格兰姆对服从有所了解。在进行**小世界**研究前的几年里，米尔格兰姆就可能因另一个巧妙的研究而大众所知，这个巧妙的研究有更令人不安的结果。20世纪60年代，早期米尔格兰姆的服从研究证明，当有权力的人（如穿着实验室外套的研究助理）要求研究被试做一些让他们不舒服的事情时，如给予（或让他们相信正在给予）另一个被试以电击，会这么做的人数量多到惊人。对于为什么有人会在即使他们不认同的情况下"服从命令"这一问题，米尔格兰姆的研究有很大的启示。

两个最近的研究已证实，社交网络中，人与人之间的平均连接数约为6或更少。

6.12.3　做一个小研究

有很多使用这些方法却不花费太多力气的方式。这个活动的目标可以是科学求证，也可以只是为了派对的乐趣。

1. 使用电子邮件

复制米尔格兰姆的研究，但利用电子邮件的便利性。在这里，问题变成了：使用电子邮箱进行联系，人与人之间的连接数为多少。电子邮件比缓慢的邮局邮件更有效，而且几乎没有成本。

当然，通过电子邮件选择被试可能更困难。很难随机选择电子邮件地址，因为没有一本大的类似电话簿的列表以供我们从中采样。此外，你发送的电子邮件可能被误认为垃圾邮件并遭忽略。顺便说一句，因为你的研究兴趣是正当的，你不必担心会违反任何互联网协议。

2. 着眼于聚会

当举办大型晚会时（如果这是一个鸡尾酒宴会，米尔格兰姆会非常喜欢，这是他最初的灵感来源），给你的宾客散发资料。给他们每人一张大的索引卡片和一支笔。每张卡片的底部都列出了参加聚会的某个宾客的名字。如果客宾不认识卡片下方列出的那个人，他应该在卡片上方签上自己的名字，并把它交给其他某个自己认识且可能认识卡片下面列出的那个人的人。

这一进程应继续下去，就像在米尔格兰姆的研究中一样，直到卡片到达列在底部的人的手里，那人便把卡片上交。晚会结束时，你可以分析数据并向你的宾客证明他们真的都相互认识。

6.12.4　只做数学计算

但是，即使没有科学研究，一个快速的数学分析也可能说服你，让你相信你和其他人之间的人数是相当小的一个数字。你知道多少人的名字？100？200？比方说，大约是100。据推测，他们每个人也都大约知道100个人的名字，所以你只通过两度分隔就已经连接到10 000个人了。（实

际上，算上你认识的一度分隔的100个人，总计是10 100人。）在你连接到一大堆的人之前不需要太多的度数，如表6-21所示。

表6-21：分隔度和相应的连接

分隔度	连　接
1	100
2	10 000
3	1 000 000
4	100 000 000
5	10 000 000 000

事实上，通过短短五度分隔，你就应该能连接100亿人，比地球总人口还多！

那么，为什么现实中，真正连接上所有人需要更多的连接数呢？问题是人与人之间的、由100个熟人构成的组群不是相互独立的。你有100个朋友，他们每个人的百人朋友圈并不是完全不同的。你比较熟悉的100个人中，有相当一部分比例的人也存在于其他人的朋友列表中。

社交网络中有很多的重叠。这种重叠实际上有助于增加你和附近的人（比如在同一国家）产生直接联系的几率。

祖父母悖论

和网络重叠类似的问题是**祖父母悖论**。你有一对父母。你的父母各自有一对父母，这样你就有4个祖父母。每个祖父母有一对父母和4个祖父母。无需往上数很多代，你就能得到数量巨大的人数。

数到40代祖父母，你需要一万亿人。这比有史以来曾经生活在地球上的所有人的总和还多。而这只发生在近千年。从哪里得到这些多出来的祖父母？也许是木星？

当然，答案是：这一过程中一定存在遗传树上的重叠。有时，一些有血缘关系的人也会结婚并生孩子。出于礼貌，我猜测他们是第二代堂兄妹，或诸如此类的关系。

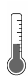
米尔格兰姆使用的小世界技术，已经在各种社交网络的研究中被认为非常有用。几度分隔的概念有一个直观的吸引力，因为它让我们觉得我们都是一个小社区的一部分。

每次我们通过某个共同朋友找到和一个陌生人的连接，这种感觉都会被加强。我不认识你，但在我自己的世界里，我是如此重要以至于我可以轻松地把自己和各种著名人物连接起来。例如，20世纪80年代初，我是堪萨斯州劳伦斯市堪萨斯大学的一名大学生，同时在美国广播公司制作的电影《明日之后》（*The Day After*）中做临时演员，这是一部广受好评的电影，讲述美国核战后的各种可能性。约翰·利特高在《明日之后》里饰演一名科学教授，他随后出现在影片《浑身是劲》（*Footloose*）中，主演就是凯文·贝肯先生！毕竟，这是一个小世界。

6

6.12.5　参阅

- 最近两个证实六度或更少的分离度的研究在《今日心理学》（*Psychology Today*）上刊登的“六度分隔”一文中有描述，作者是达比·萨克斯比（Darby Saxbe），发表于2003年11月/12月一期。
- Watts, D.J. (2003). *Six degrees.* New York: Norton.一本关于网络新科学的书，提供对我们生活的连接时代全面且引人入胜的探讨，其中包括六度分隔的概念。


HACK #73

6.13　驾驭投票循环

虽然自由选举似乎是制定政策和选举官员最公平、最明智的系统，但统计学家有时担心被政治学家称为“投票循环”的悖论，该悖论可能会导致少数群体获胜。有一种更好的方式来进行选举。

当我还是一个小小“统计学家”时，我的父母偶尔会让我对自己的事情作出选择：穿什么、吃什么、睡前读哪本故事书，等等。我注意到，有时候选择是开放式的：“你自己选择，布鲁斯。你打算什么时候去睡觉？”，有时候选择以一组选项的形式出现，我要从中作出选择：“你自己选择，布鲁斯。你想现在睡觉还是五分钟后睡觉？”

当然，第二个选择称不上是一个选择，真的。当我必须从不同的备选方案中进行选择时，我的真实想法不如我可以任意选择想要的东西时反映得那样准确。

民主就像这样。当投票选举总统、市长，或捕狗人时，我们通常要在几个备选方案之间进行选择。我们可能对任何选项都不满意，但无论如何我们都投票了（至少统计学家这么做）。但是，在离开投票间时，你是否曾经感觉那些选择在某种程度上不能完全代表你自己的真实想法？政治学家知道那种感觉。他们分析了对备选方案都不甚满意时的情况，发现在这种情况下作出选择可能会导致没人满意的结果（当然赢家除外）。

6.13.1　投票循环

选举的构成可以有多种方式。试想一下，一个选民（如一个城市居民、俱乐部会员，或大学教员）被要求表决一项政策，他有3种选择。另外，想象有3组支持者，每组支持者对其中一种意见的偏爱胜于其他两种意见。这次选举可能要求人们投票选出他们最喜欢的政策。在这个系统下，受最大的群体青睐的政策很可能会赢得最多的选票。这似乎是公平的，这也是我们最常看到的选举系统。

另一个合理的系统（至少表面上是合理的），会呈现每对对立意见，有对决选举的味道，其中A和B对比、B和C对比、C和A对比。在这种系统下，最大投票获得者的产生应该是非常公平的。但是，事实证明，这种称为投票循环的系统，很难公平地使用，因为你展示的选项顺序能决定选举结果！



选举中的投票循环和你如何安排篮球联赛的原理一样：比赛的顺序可能会影响获胜结果。

6.13.2 如何生效

以下是投票循环如何生效的一个例子。试想一下，你的童子军需要决定将部队俱乐部内部（或任何童子军集会的地方）粉刷成什么颜色。作为群体一员，你要投票给红色、白色或蓝色。不同的政治“团体”已经在你的喜欢不同颜色的同事中形成了。

有偏好红色的苹果队，有喜欢白色的大象队，有钟爱蓝色的松鸦鹰队。至于第二喜欢的颜色和最不喜欢的颜色，这些小组也有不同意见。表6-22给出了3个组和他们的政治议程。

表6-22：粉刷偏好与政治观点

组	选民百分比	第一选择	第二选择	第三选择
苹果组	20%	红	白	蓝
大象组	40%	白	蓝	红
松鸦鹰组	40%	蓝	红	白

要确定童子军的意愿，你可以举行一个两阶段的选举。第一阶段提出两个备选方案。这个阶段的获胜者随后与第三种备选方案“竞争”，挑选一个赢家。两阶段选举及其结果可能看起来如下所示。

(1) 红色还是白色？参照表6-22，红色可能会得到60%的选票，淘汰了白色。现在，获胜者去和蓝色竞争。

(2) 红色还是蓝色？在这个对决中，红色得到20%的选票，蓝色以80%的巨大支持率获胜。

因此，蓝色的油漆一定是大家的意愿！但是，这是一个悖论结果，因为只有一个组最喜欢蓝色，其人数占童子军总数的40%。同等数量的童子军最喜欢白色，而另外20%的人讨厌蓝色。决策的顺序影响了结果。让我们以不同的顺序再做一次。

(1) 红色还是蓝色？蓝色以80%的选票获胜。

(2) 蓝色还是白色？白色以60%的选票赢得这场比赛。

因为对决的顺序不同，我们得到了不同的结果。这是有趣的，让我们再做一次。也许这次我们可以安排让红色获胜。

(1) 蓝色还是白色？在这场与蓝色的对决中，白色将获得60%的选票存活下来。

(2) 白色还是红色？红色以60%的多数选票赢得这场比赛。干得好，红色。红色分明就是大家最喜欢的颜色！

3种可能的对决顺序导致3种完全不同的政策决定。

6.13.3　摆脱投票循环

如果我们把投票系统看做测量系统，那么这种做决策的对决方法具有很低的效度。在这里丢失了原本可能收集到的选民信息。但是，现在我的脑海中浮现出很多种解决循环投票问题的方法。

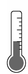
如果投票系统的设计者对选民的顺序偏好感兴趣，可以要求选民对所有候选人进行等级排序。平均等级最低者获胜（将最偏爱的一位排为等级一）。这是一个使用了所有可用信息且更公平的方法，但这将导致没有人为最终结果激动不已。



例如，多年前这样的系统导致了我家臭名昭著的决定：我们平安夜的电影是《小鬼当家》（*Home Alone*）。

另一种解决方案是给所有候选人提供一张选票，票数最多者获胜。这是最常用的系统，但当所有候选人都没获得大多数人的支持时，这种系统确实有缺点。

对于有很多候选人的选举（比如，某些市长或省长选举），经常会有一个对决，在对决中，大量的候选人被削减到一个较小的数目。这没有投票循环的缺点，因为所有的候选人在同一时间都被考虑到了。它也消除了单向投票方法的缺点，因为它增加了受大多数人支持的候选人获胜的可能性。



HACK #74

6.14　在快车道上生活（你已经在了）

通过应用几率以及对人性的认识，还有一些关于高速公路驾驶行为的事实，你可以作出更明智的变道决策。

没有什么比堵在路上更令人沮丧了，尤其当其他车移动得比你快时。虽然往快速车道上变道很有吸引力，但事实证明，你的判断可能是错误的，另一条车道也许真的不比你的车道快。

不应该变道时却决定变道，这是一个危险的决定。不仅多数的车辆碰撞事故是由错误驾驶导致的，而且美国每年发生300 000起车辆事故，尤其常发生在司机变换车道时。当然，如果你赶时间且你旁边车道上的车比你开得更快，只要你能安全变道，一个明智的司机为什么不移动到快车道上去？毕竟，正如我已经耐心地向法院机关解释了很多次一样：“好”司机不一定是更安全的司机；他只是能够尽可能快地到达目的地的司机。

问题是，最近基于计算机模拟的统计研究表明，司机通常会判断另一条车道移动得比他们的更快，即使它们实际上是以相同的速度在移动！调查研究显示，这种误解足以让大多数司机试图向其他车道变道。

6.14.1 跳过、错过以及时期

在一条繁忙的高速公路上，或交通堵塞时，我们的感官世界是由我们面前的大卡车、我们左右两侧能看到的汽车，以及堵在我们后面可怜的傻瓜构成的。判断我们的行车速度时，虽然我们有一个速度表，但最引人注目的数据往往来自我们两侧车道的汽车。（是它们超了我们还是我们超了它们？）

交通研究人员把你超过其他车辆的时刻称为*跳过*（skip），把其他车辆超过你的时刻称作*错过*（slip）。最近的研究把跳过称为*超车时期*，把错过称为*被超车时期*。相比被超车时期，司机非常喜欢超车时期，这可能不会令你感到惊讶。



一个**时期**是一段时间。交通繁忙时，司机在路上的驾驶时间本质上是由一系列很短的持续时间构成的。

除了寻找要进入的更快车道，司机还有一个目标，那就是让自己的车保持尽可能快的速度，或者至少接近目标速度（比如，可能是接近限速水平的速度）。如果感觉自己和前方车辆有一定的距离，并且自己目前尚未以目标速度移动，那么司机会加速以缩小差距。正是这些突然的加速形成了跳过（超过其他车的时期）和错过（被其他车超过的时期）。相比我们超过别的车辆，当我们被其他车辆超过时，我们可能体验到更多的时间间隔。正是这种感知上的不公正导致司机推断他们是在慢车道上，即使两个车道都一样的缓慢。

想象一下，两条并排车道以相同的平均速度移动。汽车之间间距的形成具有*随机性*；更准确地说，它们的形成具有*系统性*，但是基于一个随机的起始配置。间距形成的同时，间距也被填补；间距被填补的同时，汽车也在加速。



一条车道的平均速度可以用车辆行驶距离除以行驶时间来计算。所以，如果两条车道上的车辆均5分钟开了1000码，那么它们具有同样的每分钟200码的平均速度，或每小时6.8英里。

在拥挤的公路上，偶尔有机会供司机尝试缩小间距，但实际上，缓慢移动或不移动所花费的时间更多（相对而言）。在缓慢移动时，偶尔会有其他车道的汽车填补间隙，并超过那些暂时在慢车道上的司机，当然，这需要更多的时间。

按时期来测量时，对于任何一个司机而言，被超车的时间将比超车的时间要更多。这是因为你是在快速移动时超过其他车辆的，而在缓慢移动时被超车。图6-9展示了这种感知。

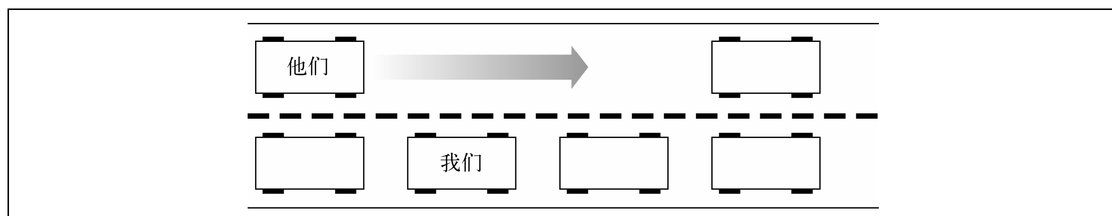


图6-9：感知被超车的时间

当发现其他车辆在加速填补间距，而我们却依然静止不动时，就形成了我们的车道移动得更慢的错觉。

6.14.2　概率与交通模式

通过进行计算机模拟来确定司机对其他车道速度感知的准确性，加拿大研究人员唐纳德·雷迈德（Donald Redelmeier）和罗伯特· 托比希赖（Robert Tobshirani）对基于正态分布特征[Hack #23]的交通模式提出了一些假设。

在拥挤的高速公路上，有几个原因（如条件、出入口等）能形成间距，为了反映这个现实，他们基于两个正态分布在移动车辆间随机分配间隔：90%的间隔为相距2米左右，加减0.1米；10%的间隔为相距100米，加减5米。在数百次模拟的最开始，车辆和间距都遵循这种随机计划被放置和创造。

研究人员创建以相同速度向同一方向行驶的两条车道的数据，他们想象出了数百个有典型加速和制动能力的车辆。它们以这样一种安全驾驶策略运行：当车道上有间距时，它们就会往前移动，但不能靠得太近。这些模拟司机不允许过于靠近其他车辆的后挡板。此外，也允许车辆改变车道，这一定令电脑控制的司机沮丧。这里没有事故。



以平均加速度和制动速度模拟他们的车辆，雷迈德和托比希赖选择典型的统计指标（在10秒内起步加速到63英里/小时的能力和在5秒从63英里/小时减速到0的能力），这恰好和本田雅阁的参数匹配。

6.14.3　作出明智的变道决策

雷迈德和托比希赖发现13%的时间，汽车要么超车要么正被超车。大多数时候，汽车行驶的速度彼此相当。任意特定的司机被超车的几率比他正在超车的几率更高，当他超车时，他超过了一堆汽车。数学运算得出，被超车的汽车和超车的汽车数量差不多。被我们的司机超过的汽车总数等于超过他的汽车数量。

在拥挤的公路上驾驶时，大部分时间里另一条车道似乎都更畅通。有一些方法来处理这样的

误解并作出更明智的（和统计安全）驾驶选择：

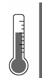
- 作为一名有逻辑的科学家，你可以通过旅程的长度来评估你的驾驶，而不是通过你是否赢得了堵车比赛来判断。如果你认为与其他车道相比，有更多的车超过了你，也没有关系。
- 记住其他车道更好这个误解，并寻找更好的方法来判断其他车道的速度。选择另一条车道上的一辆车，几分钟后对比你和它的位置。毕竟，有时会有一些更快的车道，但你不能把超过你的车辆当做速度的最佳证据。
- 在大型公路上，远离左侧或右侧将有出口的车道，因为车辆驶出或驶入道路，是减速和加速的主要原因。
- 无论驾驶还是购买汽车，都要遏制你的冲动。有趣的是，仿真结果表明，冲动的驾驶，如最小化你和另一辆车之间的距离，实际上会增加你注意到其他车辆超过你的时间。此外，更快的汽车（那些能够迅速加速的）会花更少的时间超过其他汽车，因为它们可以更快地做到这一点。所以，你的超级动力跑车可能导致你在拥挤的公路上遭受更多的挫折。

要处理以为另一条车道速度比你的车道速度快这个可能的误解，最明智的策略也许也是最简单策略，即只要不去注意它。仿真结果表明，如果你查看其他车道的时间缩减一半，那你发现汽车超过你的时间也减半。

但是我认为我们并不需要统计分析来告诉我们这一点。不要想你旁边的车，要更多地注意你后面的车。你已经遥遥领先它们了，它们有成千上万辆呢。你已经赢得了堵车比赛。

6.14.4 参阅

- Redelmeier, D.A. and Tibshirani, R.J. (1999). “Why cars in the next lane seem to go faster.” *Nature*, 401, 35. 最初的研究报告基本涵盖了本文的交通分析。
- Redelmeier, D.A. and Tibshirani, R.J. (2000). “Are those other drivers really going faster?” *Chance*, 13, 3, 8-14. 上述《自然》杂志上的文章有对结果更详尽的描述。



HACK #75

6.15 寻找新生命和新文明

搜索外星生命非常盛行。你可以使用统计抽样和概率聚焦于搜索。

对于和其他星球的生命进行通信这一科学追求，我们需要作出判断。首先，必须判断除了我们自己的星球上（我的是地球，你的是什么？）有生命，其他星球是否存在生物。二是必须确定如何以及在哪里找到它们。你可以使用统计方法来做这两个判断。

6.15.1 估计智能行星的数目

1961年，弗兰克·德雷克（Frank Drake），一位对通过读取无线电波（一大堆一直由地球反

6

射）观察宇宙遥远处感兴趣的天文学家决定估计可能存在多少种技术先进的文明。

以我们小小的银河系为中心，他最感兴趣的是确定在我们的银河系中，地球附近究竟有多少种先进文明（愿意并能够与我们交谈的行星）。德雷克给出了这个等式：

$$银河系中的文明数=(R)(N_h)(F_l)(F_i)(F_c)(L)$$

表6-23显示了德雷克方程中各缩写的含义。

表6-23：德雷克等式构成

术语	含　义
R	银河系中新恒星诞生的速率（每年）
N_h	环绕每颗可以支持生命的恒星的平均行星数
F_l	能够孕育生命的（从 N_h ）行星比例
F_i	能够孕育智慧生命（从 F_l ）的行星比例
F_c	能够发展文明（从 F_i ）的行星比例
L	文明（从 F_c ）的平均寿命（年）

这个计算公式真的只是一个概率链。预期的积极结果由所有单独的可能性相乘确定。不包含这些F变量的公式更简单而且效果也不错，将这些特定的不同部分纳入其中，能帮助科学家们确定当估算我们并不孤单的概率时，所需回答的重要问题。

6.15.2　应用德雷克方程

为了计算我们的银河系中目前存在智慧生命的真实星球数，你必须代入一些真实的数字。另外，我们知道正确答案（方程的解）一定至少为1，因为地球上有智慧生命（在这里插入你自己的笑话），而且一定不会超过恒星（可能支持生命）周围的平均行星数的250 000 000 000（银河系中恒星的数目）倍。

当第一次引入这个公式时，天文学家公认只有一个变量可以估计。那就是R，即我们的银河系每年新产生的恒星数，这个数字被认为大约是10。



如果20世纪60年代R被认为是10，我想现在我们银河系恒星的正确数量将接近2500亿+40。

1980年，天文学推广者卡尔·萨根（Carl Sagan），在他的电视节目以及同名书《宇宙》（*Cosmos*）中讨论了德雷克方程。因为我们对自己太阳系的行星了解甚少，而且，更重要的是，我们对其他太阳系（哪怕有这样的事情）的行星一无所知，所以萨根对每个值的估计以及他的最佳猜测答案都是具有推测性的，但他的回答是，在任意特定时间，银河系中都约有600万行星拥有能够

和我们沟通的技术。

根据我们今天所掌握的知识，表6-24提供了一组可以产生一个可能答案的值。这些值取自2005年10月《天体生物学杂志》（*Astrobiology Magazine*）上的一篇文章，作者是纽约大学的史蒂芬・索特博士（可能你的咖啡桌上就有一本）。在某些情况下，我从索特提供的系列值中选择了一个确切的值。

表6-24：德雷克方程的应用

项	估计	计算
R	每年10个	10
N_h	0 .01（100个恒星中有1个行星）	10 × 0.01=0.10
F_l	1（以地球为代表）	0.10 × 1=0.10
F_i	0.001（索特提出的“小分数”）	0.10 × 0.001=0.0001
F_c	0.20	0.0001 × 0.20=0.000 02
L	100 000年	0.000 02 × 100 000=2

有了这些数字，公式估计，在整个银河系中能相互通信的行星数量总共是两颗。地球是其中的一个。那另外一个是哪个？

正如萨根、索特以及其他作者指出的那样，在我们的银河系中，在任何给定时间能支持高等生命的星球数量取决于很多随意估计的因素，输入数值时，任何一个小的选择都能极大地改变结果。600万个可能的朋友和只有2个可能的朋友，这两者有着重要的区别，但两者的估计都来自合理的假设集。

当你对方程每个部分尝试不同的估计时，请注意方程的解是如何变化的。如果大多数智慧生物（比方说80%）最终会产生文明，那么可能行星的数量变为8。对于能够支持生命的恒星，如果其周围的平均行星数实际上是2（如萨根所建议），我们的8颗行星将变成1600颗行星。

索特表示，不同的合理估计能产生几千种的答案；另一方面，受限于我们自身的无线电能力，也可能产生如此少的答案，表现出统计上的不可能性，以至于我们成为成千上万的星系中唯一的先进文明。

6.15.3　寻找我们的空间密友

德雷克方程的一个可能的结果是：在我们的银河系中，只有两颗行星具有能够发送和接收无线电波的高等智慧文明。如果我们真的只有一个潜在的宇宙笔友，那么在如此多的行星里将会很难找到他或它。那么，该怎么办？

目前寻求新生命、新文明的策略是用微波接收器扫描天空。无线电信号有广泛的频谱。有一

些频谱是自然存在的，有些则属特别窄的范围，被认为只能人工创造，比如从《三人行》（*Three's Company*）电视节目或者通过雷达传输的频谱。那些寻找外星生命的人格外关注属于人工光谱的信号，他们希望发现并分离出先进文明的随机输出，当然，这也可能是任何有兴趣的观测者基于利益而故意发出的信号广播。



如果你有一批属于自己的微波监听电台，就会想把它们调至利于发现其他星球生命的频率：1.42千兆赫。任何天然的信号源都不可能在该频率发射电波。

不过，天空很辽阔，研究人员使用既有针对性又具便利性的抽样技术来确定寻找区域。搜索策略专注于满足两个条件的恒星亚群：

- 它们是与我们的太阳有共同特征的恒星；
- 它们在附近（距离地球只有100光年）。

6.15.4　数据分析

如果能发出关键生命信号的行星数量非常少（如德雷克方程中显示的数字），那么这样的样本搜索必须是非常彻底的；否则，我们可能会错过它。统计学家们把这种情形的研究归为，需要一个很大的统计检验力[Hack #8]，因为效应值是如此之小。

扫描天空时有如此多的数据被收集起来，以至于没有人，甚至没有计算机能成功分析它。你能获得帮助！SETI@home是伯克利大学的一个基础项目，安排人员定期用普通家庭或办公室电脑接收一些数据，所以当他们没有做别的事情时，他们的计算机可以对数据进行分析。SETI是的Search for Extraterrestrial Intelligence（搜寻外星文明）的缩写。该项目就像一个屏幕保护程序，你可以在http://setiathome.berkely.edu免费下载。

当你得到这些数据时，这些数据对你没有任何意义，但你的电脑会使用统计分析对信号的信息进行排序，寻找到能说明问题的、非随机的、窄带宽的、可能意味着另一个星球已经成熟到能产生《傻子派尔》（*Gomer Pyle*）或《飞跃情海》（*Melrose Place*）此类电视剧的信息。你可能是第一个发现其他星球生命的人，所以去工作吧！









看完了

如果您对本书内容有疑问，可发邮件至contact@turingbook.com，会有编辑或作译者协助答疑。也可访问图灵社区，参与本书讨论。

如果是有关电子书的建议或问题，请联系专用客服邮箱：ebook@turingbook.com。

在这里可以找到我们：

微博 @图灵教育 : 好书、活动每日播报
微博 @图灵社区 : 电子书和好文章的消息
微博 @图灵新知 : 图灵教育的科普小组
微信 图灵访谈 : ituring_interview，讲述码农精彩人生
微信 图灵教育 : turingbooks